Untangling the web of data
A critical analysis of the Archaeological Semantic Web

Thesis for the degree of Doctor of Philosophy

February 2016

Frank Lynam

Department of Classics
School of Histories and Humanities
Trinity College Dublin
University of Dublin
# Table of contents

**LIST OF TABLES** ............................................................................................................. VI

**LIST OF FIGURES** ........................................................................................................ VIII

**DECLARATION** .............................................................................................................. XIII

**SUMMARY** .................................................................................................................... XIV

**ACKNOWLEDGEMENTS** ............................................................................................... XVI

**CHAPTER 1 – THE PROBLEM WITH ARCHAEOLOGICAL INFORMATION SYSTEMS .... 1**

- Introduction ...................................................................................................................... 1
  - Thesis structure ............................................................................................................. 6

- Digital practice within the humanities and archaeology ............................................... 6
  - What are the digital humanities? .................................................................................. 6
  - The three waves model ................................................................................................. 8
  - A model of the digital humanist ................................................................................... 13
    - DHer 1 – the baseline digital humanist ..................................................................... 14
    - DHer 2 – the tool specialist ....................................................................................... 14
    - DHer 3 – the builder .................................................................................................. 15
    - DHer 4 – the project manager ................................................................................... 16
    - DHer 5 – the domain knowledge specialist ............................................................... 16
  - Digital archaeology and Archaeological Information Systems .................................... 17

- Archaeological data conceptualisation and data terminology ....................................... 21

- My ideal reader ............................................................................................................. 25

- Thesis conventions and online resources ..................................................................... 26

- Conclusions ................................................................................................................... 27

**CHAPTER 2 – PROPOSING A SOLUTION................................................................. 29**

- Introduction ................................................................................................................... 29

- A history of Linked Open Data and the Semantic Web ................................................. 29
  - The birth of the document web .................................................................................... 29
  - Web 2.0 ......................................................................................................................... 32
  - Open Access, Open Data and the politics and ethics of web access ............................ 33
  - On the problem of publishing data to the document web ........................................... 35
  - Introducing Linked Data and the Semantic Web .......................................................... 38
  - Is Semantic Web data Big Data? ................................................................................ 43
  - A note on terminology and trends .............................................................................. 45

- A technical primer on Linked Open Data and SPARQL ............................................. 47
  - Linked Open Data and RDF broken down ................................................................. 48
  - RDF triplestores and the SPARQL way ....................................................................... 56

- The Archaeological Semantic Web ............................................................................... 63
Early investigations into archaeological digital semantics ........................................64
A new millennium, a new paradigm .....................................................................66
Bringing archaeological concepts into the CRM fold ........................................68

Conclusions ...........................................................................................................70

CHAPTER 3 – DATA TRANSPARENCY AND RE-USE ........................................ 73

Introduction ..........................................................................................................73

Why the philosophical perspective? ....................................................................74
Where to begin? .....................................................................................................75
Platonic and Aristotelian ontological models ...................................................78
The European Enlightenment ............................................................................81

Knowledge Representation's appropriation of the ontology ...............................85
On the importance of Set Theory .........................................................................86
Implementing the Guarino KR ontology model for archaeology ......................86

KR ontologies in the Semantic Web ....................................................................90
Ontology authoring using OWL ..........................................................................91

Applying a Semantic Web ontology ...................................................................94
On advanced machine reasoning .......................................................................96

Semantic Web controlled vocabularies and SKOS ..........................................99

Archaeological Semantic Web ontologies .........................................................102
The CIDOC CRM ...............................................................................................102
The basic design philosophy ..............................................................................103
Archaeological extensions of the CRM .............................................................106
CRM-EH .............................................................................................................107
CRMarchaeo .......................................................................................................109

Conclusions .........................................................................................................111

CHAPTER 4 – LINKEDARC.NET ................................................................. 113

Introduction ........................................................................................................113

Linked Open Data and archaeology: the trend setters .....................................113
The Numishare framework, Nomisma and the Online Coins of the Roman Empire ..................................................114
PeriodO ................................................................................................................116
Pleiades and Pelagios ...........................................................................................119
The Getty Linked Open Data resources ..............................................................121
Europeana .............................................................................................................126
Open Context .......................................................................................................130
British Museum ..................................................................................................133
Seneschal .............................................................................................................135
CLAROS ..............................................................................................................139
Linked Open Data archaeological resource patterns .......................................140

linkedarc.net ........................................................................................................143
linkedarc.net version 1 – native RDF Turtle file system ...................................144
linkedarc.net version 2 – a hybrid triplestore using a MySQL backend ...........147
linkedarc.net version 3 – Apache JenaFuseki ....................................................149
Serving SPARQL queries ......................................................................................150
Serving URI requests ..........................................................................................151
CHAPTER 5 – POPULATING LINKEDARC.NET ........................................ 177

Introduction ........................................................................................................ 177

The Priniatikos Pyrgos Project ........................................................................ 178
Geographical, historical and research contexts ............................................. 178
The archaeological recording systems used at Priniatikos Pyrgos ............. 179
The Priniatikos Pyrgos paper record ............................................................ 181

Building an Archaeological Information System ......................................... 188
Data gathering and archaeological schools of thought ............................... 188
Assessing the user needs at Priniatikos Pyrgos ........................................... 190
The first digitization phase ............................................................................ 191

Migrating to the Archaeological Semantic Web ......................................... 193
Choosing an ontology for the mapping ......................................................... 194
Mapping the Priniatikos Pyrgos dataset to linkedarc.net ........................... 198
Messy data and data field types ................................................................. 199
Data cleaning using Open Refine ................................................................. 201
Processing number values using RDF Data Utilities ................................. 205
Processing periods using RDF Data Utilities .............................................. 208
Creating vocabulary lists out of messy data ............................................... 210
Geocoding the context and locus data ....................................................... 214
Processing image data .................................................................................. 217
Outputting the project data as RDF using RDF Refine ............................... 220
Serialising as RDF triples ........................................................................... 223
Test and deployment servers ....................................................................... 224
On inferences ................................................................................................. 226
Digitally reassembling the Priniatikos Pyrgos team .................................... 227

Conclusions ....................................................................................................... 229

CHAPTER 6 – MINING THE ARCHAEOLOGICAL SEMANTIC WEB ........ 233

Introduction ....................................................................................................... 233

Returning to Priniatikos Pyrgos ................................................................. 234
Asking questions of Priniatikos Pyrgos on linkedarc.net ............................ 234

Visualising the stratigraphy of Priniatikos Pyrgos .................................... 240
Context ........................................................................................................... 240
Research question ......................................................................................... 242
Methodology ................................................................................................. 242
Research findings ........................................................................................ 254
The loomweights of Priniatikos Pyrgos .......................................................... 255
Research question .......................................................................................... 256
Context ........................................................................................................... 256
The loomweight form and their use within the production of textiles .......... 258
Methodology .................................................................................................. 261
Recording the Priniatikos Pyrgos loomweights .......................................... 261
Querying the loomweight data using SPARQL ............................................ 265
The PPTextiles web app ............................................................................... 266
Research findings ......................................................................................... 272

Interpreting the British Museum cuneiform collection ............................. 274
Context ........................................................................................................... 274
Research question ........................................................................................ 277
Methodology ................................................................................................ 277
Getting a sense of the British Museum’s cuneiform Linked Open Data ....... 277
Cleaning the data .......................................................................................... 282
Mapping the collection’s cuneiform ............................................................ 283
Focussing in on the contributors ................................................................. 284
The cuneiform transliterations and translations .......................................... 287
Research findings ......................................................................................... 292

Conclusions .................................................................................................. 293

CHAPTER 7 – A CRITICAL ANALYSIS OF THE ARCHAEOLOGICAL SEMANTIC WEB 297
Introduction .................................................................................................... 297

Comparing the Semantic Web to the alternatives .................................... 298
FileMaker Pro and offline Archaeological Information Systems ............... 300
MySQL and the closed online Archaeological Information System ......... 302
NoSQL, and Big Data Archaeological Information Systems .................... 307
Alternatives in review ..................................................................................... 308

A sociological perspective of the Archaeological Semantic Web .......... 309
Existing scholarship on the social Archaeological Semantic Web .......... 311
The linkedarc.net ‘Attitudes to Digital Data Sharing within Archaeology’ survey results .......................................................... 314
Finding patterns in the ARIADNE and linkedarc.net survey results ...... 319
Semantic Web workshops in review ............................................................... 319
The Archaeological Semantic Web user model ......................................... 325
ASWer 1 – baseline Archaeological Semantic Web user ......................... 327
ASWer 2 – advanced Archaeological Semantic Web tool user ................. 328
ASWer 3 – Archaeological Semantic Web maker ...................................... 329

The epistemological implications of the ASW user model .................... 330
What is classical epistemology? ................................................................. 330
What is digital epistemology? ................................................................. 332
Accommodating the ASWer model within digital epistemology .......... 336
On the possibilities and challenges of the Archaeological Semantic Web .... 337
Data linking and dynamic knowledge systems ........................................ 338
On finding data on the Archaeological Semantic Web and data myopia .... 339
Epistemological determinism ...................................................................... 340

Accommodating different archaeological philosophies within the Archaeological Semantic Web ......................................................... 342

Conclusions .................................................................................................. 345
## APPENDICES

### Appendix A – Priniatikos Pyrgos RDF mappings using Open Refine and RDF Refine

- Mapping the image data.......................................................... 369
- Mapping the catalogued ceramic and object data........................ 371
- Mapping the context data......................................................... 373
- Mapping the sub-context data.................................................... 376
- Mapping the locus data........................................................... 378
- Mapping the pail data............................................................. 379
- Mapping the period data.......................................................... 380
- Mapping the controlled vocabulary data.................................... 381
- Mapping the team data........................................................... 382

### Appendix B – Attitudes to digital data sharing within archaeology survey

- The basics................................................................................. 384
- Your thoughts on digital archaeology........................................ 385
- On digital data sharing............................................................. 386

### Appendix C – Semantic Web specialised user workshop

- Participant survey................................................................. 389
- Exercise 1 – Did Frank Lynam excavate a lot of pottery while working at Priniatikos Pyrgos?........ 390
  - Objective............................................................................. 390
  - Guidelines........................................................................... 390
  - Review of exercise 1......................................................... 390
- Exercise 2 - Create a timeline of the Priniatikos Pyrgos excavation.............................................. 391
  - Objective............................................................................. 391
  - Guidelines........................................................................... 391
  - Resources............................................................................ 392
  - Review of exercise 2......................................................... 392

### Appendix D – Glossary of terms

-.................................................................................................. 394

## BIBLIOGRAPHY

................................................................. 401
List of tables

Table 1: an N-Triple example ........................................................................................................... 52
Table 2: an example RDF Turtle serialisation .............................................................................. 54
Table 3: an example RDF Turtle serialisation .............................................................................. 57
Table 4: a simple SPARQL SELECT query .................................................................................. 60
Table 5: a SPARQL SELECT query with constants ..................................................................... 61
Table 6: chaining criteria within a SPARQL SELECT query ......................................................... 62
Table 7: specifying an archaeological conceptualisation using natural language .................... 87
Table 8: specifying an archaeological conceptualisation using natural language .................... 88
Table 9: an OWL ontology for an archaeological site ................................................................. 92
Table 10: instantiating the archaeological site ontology ............................................................. 95
Table 11: a SPARQL query addressed at the archaeological site RDF dataset ......................... 95
Table 12: combining two ontologies using OWL ......................................................................... 98
Table 13: a Getty AAT SKOS declaration for its silver metal concept ........................................ 101
Table 14: matching free text data to thesaurus terms by the Seneschal project (Binding 2014b) ............................................................................................................................................ 137
Table 15: the linkedarc.net SPARQL endpoint parameter list ........................................................ 150
Table 16: a SPARQL SELECT query to get the first 10 RDF triples in the triplestore. 150
Table 17: linkedarc.net v3.0.0 web service methods .................................................................. 154
Table 18: an example linkedarc.net display schema file ............................................................. 164
Table 19: using the GREP replace function .............................................................................. 203
Table 20: cleaning dates using GREP .......................................................................................... 204
Table 21: adding inferencing support to a Priniatikos Pyrgos image resource ......................... 226
Table 22: a SPARQL SELECT query used to generate a list of the Priniatikos Pyrgos team members ........................................................................................................................................... 228
Table 23: a SPARQL SELECT query asking for all contexts and loci containing human bones ........................................................................................................................................ 237
Table 24: a SPARQL SELECT query asking for contexts/loci with human bone and carbon remains ........................................................................................................................................... 239
Table 25: retrieving the stratigraphic sequence of Priniatikos Pyrgos Trench 2 using SPARQL ........................................................................................................................................... 244
Table 26: finding the more materially significant contexts in Priniatikos Pyrgos Trench 2 using SPARQL ........................................................................................................................................... 253
Table 27: Priniatikos Pyrgos loomweight type groupings and their artefact counts .................... 262
Table 28: finding Priniatikos Pyrgos’s discoid loomweights using SPARQL ............................... 265
Table 29: finding the Classical loomweights at Priniatikos Pyrgos using SPARQL ................. 266
Table 30: locating the Greek loomweights in the British Museum using SPARQL ................. 273
Table 31: counting up all of the artefacts in the British Museum using SPARQL ..................... 278
Table 32: retrieving the first 100 British Museum artefacts with visual representations using SPARQL ........................................................................................................................................ 278
Table 33: retrieving the first 100 British Museum artefacts with inscriptions using SPARQL ........................................................................................................................................ 280
Table 34: retrieving the first 100 British Museum artefacts with cuneiform inscriptions using SPARQL ........................................................................................................................................ 281
Table 35: retrieving contextual information about the British Museum cuneiform inscriptions using SPARQL ........................................................................................................................................ 282
Table 36: getting the birth and death dates of Dr Hall from the British Museum using SPARQL ........................................................................................................................................ 286
Table 37: retrieving images of people from DBpedia using SPARQL..............................287
Table 38: retrieving the transliterations of the British Museum’s cuneiform inscriptions using SPARQL ..............................................................................................................288
Table 39: results of running the cuneiform notes through the Stanford Named Entity Recognizer tool..................................................................................................................................................290
Table 40: a simple SQL statement example ..........................................................................................................................303
Table 41: posing a multi-table query to a SQL database.............................................................................................................306
Table 42: Semantic Web workshop 3 – exercise 1 instructions.................................................................................................324
Table 43: Semantic Web workshop 3 – exercise 2 instructions.................................................................................................325
List of figures

Figure 1: the growth of internet users (created with Plotly, after internet Live Stats (2015)) .......................................................... 33
Figure 2: the Linked Open Data cloud as of 30 August 2014 .......................................................... 41
Figure 3: the frequency of Google searches related to Linked Open Data and the Semantic Web since 2004 (Google 2015b) .......................................................... 45
Figure 4: Spivack’s visual interpretation of the web’s development as iterations (Spivack 2007) ........................................................................ 46
Figure 5: the frequency of Google searches for ‘Web 2.0’ and ‘Web 3.0’ since 2004 (Google 2015b) ........................................................................ 46
Figure 6: an RDF triple’s structure ........................................................................ 50
Figure 7: the Context 101 example as an RDF triple ........................................................................ 50
Figure 8: the Context 101 example as an RDF triple using URIs ........................................................................ 51
Figure 9: a more complicated meaning rendered using RDF triples ........................................................................ 51
Figure 10: the Context 101 RDF triple example including a literal ........................................................................ 52
Figure 11: different RDF serialisations all encoding the same meaning ........................................................................ 53
Figure 12: the updated Semantic Web ‘layer cake’ (Wikipedia 2015) ........................................................................ 58
Figure 13: the Apache Jena Fuseki SPARQL query interface ........................................................................ 59
Figure 14: the results of running the first SPARQL query against the Context 101 dataset ........................................................................ 61
Figure 15: the results of running the second SPARQL query against the Context 101 dataset ........................................................................ 61
Figure 16: the results of running the third SPARQL query against the Context 101 dataset ........................................................................ 62
Figure 17: the RDF graph crawled by the third SPARQL query........................................................................ 62
Figure 18: a histogram showing the dates of Linked Open Data-related resources selected by LODLAM ........................................................................ 71
Figure 19: a conceptualisation of the archaeological site knowledge domain ........................................................................ 87
Figure 20: the journey of real phenomena from reality to ontologies (after Guarino et al. (2009 fig. 2)) ........................................................................ 89
Figure 21: defining an ontology relation using OWL ........................................................................ 93
Figure 22: the results of running the aforementioned SPARQL query against the archaeological site dataset ........................................................................ 96
Figure 23: interrogating multiple ontologies using logical reasoning ........................................................................ 97
Figure 24: Ramon Llull’s Tree of Science (Lima 2011 fig. 11) ........................................................................ 99
Figure 25: relations between the core CRM classes (Doerr 2003 fig. 3) .......................................................... 104
Figure 26: a graph modelling the signing of the Yalta Agreement using the CRM (Sinclair 2015) ........................................................................ 105
Figure 27: a graph describing the structure of the CRM-EH model (Binding 2011) ........................................................................ 107
Figure 28: an example showing the mapping of the CRMarchaeo classes onto archaeological phenomena (Cripps et al. 2014 fig. 4) ........................................................................ 110
Figure 29: a distribution map showing the find spots and mint locations for the OCRE coins ........................................................................ 115
Figure 30: the Nomisma record for Carian mint, Apollonia Salbace, represented as RDF Turtle ........................................................................ 115
Figure 31: period assertions related to the ‘Iron Age’ concept for the Levant region (Rabinowitz 2014 fig. 3) ........................................................................ 118
Figure 32: the Pleiades data model ........................................................................ 120
Figure 33: Arthur Evans as represented by the Getty ULAN data service ........................................................................ 122
Figure 34: the semantics of the Getty Vocabulary Program ontology ........................................................................ 124
Figure 35: an HTML view of the Getty Kamares concept (Harpring et al. 2015) .................................................. 125
Figure 36: a view of the Europeana digital object related to the site of Newgrange, Ireland (Europeana Project 2015c) ........................................................................................................ 126
Figure 37: an object-centric modelling of Da Vinci’s Mona Lisa by Europeana (Isaac 2013 fig. 7) .................................................................................................................. 128
Figure 38: an event-centric modelling of Da Vinci’s Mona Lisa by Europeana (Isaac 2013 fig. 10) .................................................................................................................. 128
Figure 39: a graph view of the Europeana Newgrange resource .................................................................................. 129
Figure 40: a map showing the locations of the projects published by Open Context .... 131
Figure 41: a trench record from the Kenan Tepe project on the Open Context system 132
Figure 42: an HTML view of 26th Dynasty Egyptian sarcophagus found using the British Museum’s SPARQL endpoint ........................................................................................................ 134
Figure 43: an English Heritage crop mark concept as viewed on the Seneschal system 138
Figure 44: a coin record displayed on the CLAROS system .................................................................................. 140
Figure 45: archaeological Linked Open Data service providers categorised by type .... 141
Figure 46: linkedarc.net ver1 server framework .................................................................................................. 144
Figure 47: the linkedarc.net data hierarchy ........................................................................................................... 145
Figure 48: linkedarc.net ver2 server framework .................................................................................................. 147
Figure 49: linkedarc.net ver2 bridging RDF and MySQL ....................................................................................... 148
Figure 50: linkedarc.net ver3 server framework .................................................................................................. 149
Figure 51: linkedarc.net ver3 SPARQL bridge ........................................................................................................ 149
Figure 52: linkedarc.net ver3.0.0 processing a URI request .................................................................................. 151
Figure 53: the processing of a web service call by linkedarc.net v3.0.0 ................................................................. 152
Figure 54: a web app linkedarc.net client ............................................................................................................... 154
Figure 55: an iOS linkedarc.net client .................................................................................................................. 155
Figure 56: a data mining linkedarc.net client ......................................................................................................... 155
Figure 57: the conceptual difference between client-side and server-side web apps ............................................. 157
Figure 58: linkedarc.net making an AJAX web service call .................................................................................. 158
Figure 59: the linkedarc.net web app’s responsive layout ................................................................................. 159
Figure 60: linkedarc.net web app header menu ................................................................................................ 160
Figure 61: the linkedarc.net web app home page .................................................................................................. 161
Figure 62: navigating the linkedarc.net web app’s home page map .................................................................... 161
Figure 63: the linkedarc.net web app project page ............................................................................................. 161
Figure 64: the linkedarc.net web app datatable page .......................................................................................... 163
Figure 65: resolving a linkedarc.net display schema placeholder ....................................................................... 165
Figure 66: the linkedarc.net record page ............................................................................................................. 167
Figure 67: placement of the display schema HTML in a linkedarc.net web app record page .......................................................... 167
Figure 68: RDF graph data with multiple objects linked via the same predicate .................................................. 168
Figure 69: processing a linkedarc.net display schema file ................................................................................... 168
Figure 70: SKOS concepts as object values in the linkedarc.net web app .................................................................. 169
Figure 71: linking to similar Semantic Web resources using SKOS concepts ......................................................... 170
Figure 72: using the linkedarc.net web app Workbench ..................................................................................... 170
Figure 73: searching within the linkedarc.net web app’s record page .................................................................. 171
Figure 74: the linkedarc.net faceted search interface ............................................................................................ 173
Figure 75: adding criteria to the linkedarc.net search interface ............................................................................. 173
Figure 76: linkedarc.net search results ................................................................................................................. 174
Figure 77: the linkedarc.net SPARQL query editor ............................................................................................ 175
Figure 78: the linkedarc.net SPARQL query editor options ................................................................................ 176
Figure 79: the location of Priniatikos Pyrgos (after (Molloy et al. 2014 fig. 1)) ................................................. 178
Figure 80: a cross-section describing the Locus-Pail system of excavation and recording .................................................................180
Figure 81: the types of paper record created by the Priniatikos Pyrgos project.................................................................181
Figure 82: the Priniatikos Pyrgos C676 context sheet........................................................................................................182
Figure 83: the Priniatikos Pyrgos 002-1038 sub-context sheet .......................................................................................184
Figure 84: the Priniatikos Pyrgos sample #3 sheet.............................................................................................................185
Figure 85: a Priniatikos Pyrgos section drawing including C128, C792, C852...............................................................186
Figure 86: the Priniatikos Pyrgos 14-7852 catalogue sheet ..............................................................................................187
Figure 87: a bubble chart showing the relative sizes of the various Priniatikos Pyrgos FileMaker tables .................................................................192
Figure 88: the linkedarc.net system ontology ......................................................................................................................195
Figure 89: the Priniatikos Pyrgos ontology’s hierarchy of classes .....................................................................................196
Figure 90: mapping data from FileMaker to linkedarc.net ..............................................................................................199
Figure 91: the data structure spectrum ............................................................................................................................200
Figure 92: messy data in Open Refine ............................................................................................................................................202
Figure 93: the results of processing the 'Excavator(s)' field in the contexts table .................................................................203
Figure 94: converting messy date values into RDF date format using Open Refine ..........................................................204
Figure 95: the distribution of fine-grade pottery weights across all Priniatikos Pyrgos pail records ......................................................................................................................................................205
Figure 96: the RDF Data Utilities application displaying a summary of the contents of a number column ......................................................................................................................................................206
Figure 97: the results of a RDF Data Utilities number unit alignment .................................................................................................................................207
Figure 98: a sample RDF Data Utilities value mapping plan for periods ...............................................................................................209
Figure 99: the hierarchy of Priniatikos Pyrgos periods ........................................................................................................210
Figure 100: Priniatikos Pyrgos FileMaker fields mapping onto vocabulary values ..........................................................211
Figure 101: the Priniatikos Pyrgos custom vocabulary value groupings ................................................................................212
Figure 102: the Open Refine vocabulary extraction method ..............................................................................................213
Figure 103: the Priniatikos Pyrgos Trench 2 contexts as represented by QGIS ........................................................................215
Figure 104: a screenshot showing the KML data for context 1 in Open Refine .................................................................................................................................216
Figure 105: a photograph with the identifier 05-1006-01 of the object 05-1006, a pyramidal loomweight .................................................................................................................................218
Figure 106: a FileMaker images table record ....................................................................................................................218
Figure 107: the Priniatikos Pyrgos images hosted by Flickr ....................................................................................................219
Figure 108: the FileMaker images table export workflow ........................................................................................................220
Figure 109: the merged images data in Open Refine ........................................................................................................220
Figure 110: the RDF 'skeleton' for the Priniatikos Pyrgos images table .................................................................................221
Figure 111: the mapping of Open Refine table data to an RDF Turtle serialisation ...........................................................222
Figure 112: the outputted RDF Turtle file for the contexts table .........................................................................................223
Figure 113: a visualisation of the network created by 100,000 of the 5,985,261 triples in the Priniatikos Pyrgos dataset ......................................................................................................................................................224
Figure 114: locating all of the human bones in the Priniatikos Pyrgos dataset ........................................................................236
Figure 115: the linkedarc.net web app search workflow ........................................................................................................236
Figure 116: the contexts and locus entities containing human bone .....................................................................................238
Figure 117: the resultant Gephi nodes datatable after running a linkedarc.net SPARQL query using the Semantic Web Import plugin ......................................................................................................................................................246
Figure 118: the corresponding Gephi edges datatable ........................................................................................................246
Figure 119: a Gephi generated network graph describing the stratigraphic relationships between the Priniatikos Pyrgos Trench 2 contexts ......................................................................................................................................................247
Figure 120: periods and notes included in a Harris Matrix (Harris 1989 fig. 51) ........................................................................248
Figure 121: applying bands of colour, categorised by period, to the graph shown in Figure 119 ......................................................................................................................................................249
Figure 122: a Gephi generated network graph describing the stratigraphic relationships between the Priniatikos Pyrgos Trench 2 contexts (colour-coded by period) ...........................................249
Figure 123: the Trench 2 stratigraphic network with Early Minoan contexts brought to the fore ........................................................................................................250
Figure 124: a close-up of the network shown in Figure 122 ..................................................................................251
Figure 125: a close-up of the Priniatikos Pyrgos Trench 2 stratigraphy including relative ceramic weights ........................................................................................................253
Figure 126: 1st Millennium early Han Dynasty bronze figurine showing a weaver (Barber 1991 fig. 3.1) ..................................................................................258
Figure 127: a loom reconstruction, Gordion Terrace Building (Burke 2007 fig. 10.6) ........................................259
Figure 128: a distribution of loomweight types found at Priniatikos Pyrgos ..............................................262
Figure 129: the range of loomweight types identified at Priniatikos Pyrgos ..............................................263
Figure 130: screenshot of a Priniatikos Pyrgos biconical loomweight taken in Autodesk 123D Catch ..................................................263
Figure 131: an RDF representing the loomweight 05-1035 ................................................................................264
Figure 132: the PPTextiles web app ......................................................................................................................267
Figure 133: a heatmap showing the distribution of all Priniatikos Pyrgos loomweight finds in the PPTextiles web app ........................................................................................................267
Figure 134: filtering the loomweight distribution by type ..................................................................................268
Figure 135: a PPTextiles query response report .................................................................................................268
Figure 136: the distribution of discoid loomweight finds at Priniatikos Pyrgos ..............................................269
Figure 137: the distribution of the Bronze Age loomweights found at Priniatikos Pyrgos ..................................................270
Figure 138: querying a Priniatikos Pyrgos loomweight find spot ..................................................................270
Figure 139: the distribution of the Archaic and Classical loomweights found at Priniatikos Pyrgos ........................................................................................................................................271
Figure 140: the distribution of the Hellenistic and Roman loomweights found at Priniatikos Pyrgos ..................................................................................................................................................271
Figure 141: the CIDOC CRM's guidelines on the use of the ecrm:P65_shows_visual_item predicate ..............................................278
Figure 142: a portion of the CIDOC CRM's class hierarchy ..................................................................................279
Figure 143: the British Museum's representation of the cartouche of Malonaqan inscription RDF resource ........................................................................................................................................281
Figure 144: the results of the cuneiform SPARQL query displayed in Excel ..............................................282
Figure 145: map showing the distribution of the find spots for the British Museum's cuneiform inscriptions ........................................................................................................................................283
Figure 146: a popup window providing more information about the site of Nineveh on the cuneiform distribution map ........................................................................................................................................283
Figure 147: the British Museum's cuneiform inscription collection by discovering nation ........................................................................................................................................284
Figure 148: the top ten contributors to the British Museum's cuneiform inscription collection ........................................................................................................................................285
Figure 149: a timeline of the British Museum's cuneiform inscription contributors ..............................................286
Figure 150: a Wordle visualisation showing the relative frequencies of the transliterated words in the British Museum's cuneiform collection ........................................................................................................289
Figure 151: a Wordle visualisation showing the relative frequencies of words in the cuneiform notes identified by the Stanford NER application as being of the PERSON class ........................................................................................................................................290
Figure 152: a Wordle visualisation showing the relative frequencies of words in the cuneiform notes identified by the Stanford NER application as being of the LOCATION class ........................................................................................................................................291
Figure 153: the Semantic Web broken down into processing layers.................................299
Figure 154: a screenshot of the FileMaker Pro ver. 10 application showing a custom layout.................................................................300
Figure 155: the nature of the MySQL data infrastructure .............................................304
Figure 156: an example MySQL archaeological database ............................................305
Figure 157: alternative information systems rated by Berners-Lee's 5 stars of Open Data .............................................................................308
Figure 158: outcomes of the ARIADNE D2.1 report on user needs (ARIADNE 2015a p. 155).................................................................................................310
Figure 159: the ‘Russian Doll’ model of archaeological data sharing.............................311
Figure 160: the location of the ‘Attitudes to Digital Data Sharing within Archaeology’ survey respondents..............................................................................315
Figure 161: ‘Attitudes to Digital Data Sharing within Archaeology’ survey demographics..................................................................................316
Figure 162: ‘Attitudes to Digital Data Sharing within Archaeology’ relevance and familiarity.........................................................................................316
Figure 163: ‘Attitudes to Digital Data Sharing within Archaeology’ data sharing responses ..........................................................................................318
Figure 164: the three Archaeological Semantic Web user types ....................................327
Figure 165: knowledge pyramid (Kitchin 2014 fig. 1.1) ..............................................330
Figure 166: a representation of Liu's 'flat' model for digital epistemology......................336
Figure 167: missing potentially relevant data using the current Archaeological Semantic Web search model........................................................................340
Figure 168: modelling multivocality using RDF triples and the CIDOC CRM..................344
Figure 169: an example of SPARQL's inability to query across datasets – note the inaccessible data........................................................................362
Figure 170: linkedarc.net displaying the results of the question ‘Where is Context 1?’ ...(364
Figure 171: the LodLive tool representing the associations of the Priniatikos Pyrgos Context 1 resource................................................................................365
Figure 172: LodLive displaying associated image resources...........................................366
Figure 173: LodLive locating resources on a map..........................................................366
Figure 174: the organisation of the RDF classes employed to map the Priniatikos Pyrgos images table...........................................................................369
Figure 175: the organisation of the RDF classes employed to map the Priniatikos Pyrgos catalogue tables....................................................................371
Figure 176: the organisation of the RDF classes employed to map the Priniatikos Pyrgos contexts table......................................................................374
Figure 177: the organisation of the RDF classes employed to map the Priniatikos Pyrgos sub-contexts table...................................................................376
Figure 178: the organisation of the RDF classes employed to map the Priniatikos Pyrgos loci table.................................................................................378
Figure 179: the organisation of the RDF classes employed to map the Priniatikos Pyrgos pails table.................................................................................379
Figure 180: the organisation of the RDF classes employed to map the Priniatikos Pyrgos periods table.........................................................................380
Figure 181: the organisation of the RDF classes employed to map the Priniatikos Pyrgos vocabulary table.....................................................................381
Figure 182: the organisation of the RDF classes employed to map the Priniatikos Pyrgos actors table........................................................................382
Figure 183: a timeline showing the history of excavation at Priniatikos Pyrgos..............391
Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University’s open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Frank Lynam
Summary

This thesis begins with the idea that the archaeological data exchange system is fundamentally broken. Archaeologists are now producing more raw data than has ever been the case in the history of the discipline. Not only are archaeological datasets growing in size but their form is expanding in variety as well. It is now not uncommon for even the most modest archaeological project to produce gigabytes of digital data as an output of its labour, and an evolving set of digital practices is at the beating heart of this invention. 3D models, digital field photography, geophysical survey data and digital databases are redefining the very nature of the archaeological record, and, increasingly, we are seeing this digital creation move closer to the trowel’s edge.

From the outside looking in, these changes predict an exciting future for the field. All of this new source material must surely lead archaeology into a stimulating new era of discovery. I argue in this work, however, that the current reality strays far from this ideal. The central problem is that very little of this valuable digital data is allowed to flow freely around the epistemological channels of the discipline. This results in a number of negative outcomes, from the duplication of effort to the creation of badly structured datasets. Ultimately, this creates a propensity within the field towards ‘small question’ research.

This thesis argues that archaeology has yet to take the final step on its road to digital adoption. If the intention is really to establish a new paradigm for its research based on this digital opportunity, then it must be prepared to open up its datasets and to establish a more networked epistemological environment. To realise this dream, I advocate the adoption of Semantic Web methodologies in all aspects of the field’s work that are applicable, in an effort to establish an Archaeological Semantic Web. I argue that, despite its shortcomings, the Archaeological Semantic Web represents a workable middle ground between archaeological knowledge liberalism on the one hand and overly prescriptive standards-based epistemology on the other.

I approach this subject from a digital humanities perspective, and in that context, Chapter 1 provides a history of this field. Chapter 2 continues in this historical vein by recounting the story of the web from its document web beginnings to the proposal and refinement of the Semantic Web model by Tim Berners-Lee and others. This history
concludes with an account of how the archaeological field has been influenced by Semantic Web practices over the course of the last two decades. Chapter 2 also introduces the technologies that provide the baseline functionality for the Semantic Web: RDF and SPARQL. Chapter 3 turns its attention to ontologies, which serve to inject meaning into Semantic Web data. Chapter 3 begins by charting the course of the ontological idea from its Greek Presocratic origins to its role as a tool of Knowledge Representation with the advent of computers in the 20th century. Chapter 3 concludes with an introduction to the CIDOC CRM and its archaeological extensions, which are referenced widely throughout the remainder of the work.

Chapters 4-6 present an account of the Archaeological Semantic Web in practice. The first of these chapters presents linkedarc.net, which is the chief practical contribution of this research project. linkedarc.net is an RDF-based system designed to allow archaeological projects publish their content to the Archaeological Semantic Web. Chapter 5 recounts how the data of the Priniatikos Pyrgos Project was imported into the linkedarc.net system. Chapter 6 then describes three separate case studies that illustrate how the Archaeological Semantic Web can be used as a research tool. This emphasis on the consumer side of the model is another significant contribution of this work.

Chapter 7 adopts a more critical perspective of the Archaeological Semantic Web. It begins by comparing the model against the leading alternatives, which might conceivably be employed to address the same underlying challenge. None of these systems were found to be entirely satisfactory. Chapter 7 goes on to consider the Archaeological Semantic Web as a sociological concern. Its analysis of the findings of sociological research on the subjects of digital method adoption and data sharing within archaeology and of a series of workshops undertaken as part of this project culminate in the formulation of a tripartite model for Archaeological Semantic Web use. The ASWer model outlines the requirements, work practices and output types that characterise engagement with the Archaeological Semantic Web. Chapter 7 lastly addresses the epistemological implications of the ASWer model. It concludes that the Archaeological Semantic Web is mostly consistent with a classical epistemological view. Chapter 8 completes the thesis by looking to the future of the field. It makes a number of suggestions as to how the Archaeological Semantic Web model might be improved upon to increase its popularity among the research community.
Acknowledgements

It is one of the great ironies of the humanities doctoral thesis that only one name can ever be placed alongside the final draft when in reality it is nothing if not a work of great collaboration. Some of this collaboration takes the form of traditional academic discourse and debate, while still more of it is of the pastoral variety. At this remove, I am still too close to say which dominates the other but I am certain that both are absolutely necessary, if the effort is to have a chance of a favourable outcome.

I want to begin by thanking my supervisor, Dr Christine Morris, who has over the last four years always made herself available to thrash out whatever idea that happened to be occupying my mind at the time. Through those numerous conversations I came to mould, sculpt, break down and build anew those arguments that now form the basis of the text that you hold in your hands. Thank you Christine and to Martine Cuypers, Hugh Denard, Winifred Ryan, Hazel Dodge, Anna Chahoud, Brian McGing, Ashley Clements and to all of the other staff of the Department of Classics at Trinity College Dublin. I have yet to meet a more welcoming and supportive department than Trinity Classics. It is truly a testament to all that is good about the institution.

I was also a student of the Digital Arts and Humanities programme during the writing of this thesis. Thank you to all involved in the project. It is an extremely worthwhile and forward-thinking endeavour, and I hope that my colleagues and I are the first of a long line of doctoral students to exit its gates. One of the most pleasing aspects of working with the DAH has been the opportunity that it gave me to work with some amazingly talented people. I enjoyed the many chats about subjects that I once knew very little of and now perhaps know something more. Thank you Bilu, Dara, Emily, Roman, Catherine, Karolina, Radek, Toma and all the other DAH researchers.

When you finish reading this thesis, you will realise just how important the Priniatikos Pyrgos Project has been to its creation. I want to acknowledge the enormous support that all of the team at Priniatikos Pyrgos, but particularly Barry Molloy, Jo Day and Vera Klontza-Jaklova, have given me ever since I first joined its ranks in the summer of 2007. I have learnt a huge amount about life and archaeology from all of you, and I will be forever grateful for that gift.
As this project neared its end, I was blessed to be able to call upon a small army of proofreaders to help highlight my errors. Thank you Harry, Caroline, Una, Henry, Ron, Lena, and of course Christine for suppressing any natural ambivalence you might have held towards the Archaeological Semantic Web. Ron, as one, who has navigated these byways before me, your guidance and humour helped pull me out of many a blind alley of investigation.

I want to end by acknowledging the constant support of those closest to me: my father Harry, mother Caroline, siblings Una, Henry, Garrett, Stephen and all the extended family. And in time honoured fashion of saving the best till last, thank you Lena for being you and for making this last lap a strangely enjoyable experience.

I dedicate this project to the memory of Laurence Berry who so tragically left us during the final stages of its writing. Too young, brilliant and too soon.
Chapter 1 – The problem with Archaeological Information Systems

Outlining the project research questions and intellectual precedents

‘Introspective Digital Archaeology, therefore, seeks to understand the nature of the computational turn in archaeology and its effect on every stage of knowledge creation – from the theories we develop and use to the recognition of archaeological features on and in the ground, from the definition and capture of archaeological data through to the methods of structuring and recording those data, from their manipulation and analysis through to the presentation and synthesis of those data, and, ultimately, through to the construction, management, and publication of the resultant knowledge.’ (Huggett 2014)

‘Today, we need to understand the process of technical evolution given that we are experiencing the deep opacity of contemporary technics; we do not immediately understand what is being played out in technics, nor what is being profoundly transformed therein, even though we unceasingly have to make decisions regarding technics, the consequences of which are felt to escape us more and more.’ (Stiegler 1998 p. 21)

Introduction

I invite you to think about the following research scenario. If you are an archaeologist, the details should sound familiar and if you are not, you may be reminded of an analogous practice within your own field. Let us envisage a situation in which you are a researcher that is interested in finding out about a particular type of ceramic vessel. For convenience, let us call it a Type IIA vessel. For the experiment, it does not really matter what the nature of the physical appearance of this vessel type is, what the vessel was used for or indeed when and where it was used. These details are, of course, vitally important in any real archaeological investigation but for now they are incidental to our exercise. The challenge that I want you to think about is conceptually very simple. You must find ways of retrieving all of the information relating to these Type IIA vessels, which have been recorded by archaeologists working at two separate archaeological sites (Site A and Site B).
Chapter 1 - The problem with Archaeological Information Systems

The question that I want you to address is how would you approach this problem? Or more specifically, how would you go about retrieving this information in the 21st century? My suspicion is that you would first want to check if there existed a digital correlate of these vessels. Thankfully, archaeologists have been creating digital representations of the materials that they study for a number of years now and, as such, it is highly likely that at some stage in the lives of these two archaeological projects digital records concerning these Type IIA vessels were created. The problem now becomes more refined. We now want to know how best to find the digital records for these Site A and Site B Type IIA vessels.

The most obvious place to begin this search would be on the internet, which, as we know, is the largest store of digital information currently in existence. Luckily, technologists have been working on the problem of sorting through the vast stores of digital data that make up the internet for quite a while and they have created a number of services, which allow users of the internet to pinpoint digital resources that are relevant to their search criteria. You could use Google, the most popular of these services, to search for the names of the sites and if you are lucky, you might find a website or a series of websites dedicated to the archaeology of Site A and Site B. If these websites are up to date (never a certainty), you could then continue your investigation in confidence by browsing or, if available, by using a faceted search facility to drill down within each site’s individual dataset for details about the Type IIA vessels. If all goes well, you will find some relevant information, which you can note and set aside for later analysis.

This approach would work satisfactorily in cases where the websites contained only a few relevant records but it becomes more convoluted and unwieldy once this record figure begins to grow. In certain circumstances, it might be possible to download a file, which contains all or a portion of the total site data. This information could then be parsed in order to make sense of it and to filter out all of the irrelevant ‘noise’. If the file were structured (for example, as an Excel or CSV file), the effort demanded by this filtering process would be attenuated but this would increase dramatically were the information to be returned in an unstructured format (as a text document, for example). And then you would need to take into account the possibility (or, more likely, the probability) that the data of either site may be represented in different ways. In this case, you would need to ascertain how to best merge the different formats so as to allow for their combined
analysis. And at the end of accomplishing all of this, you would most likely also need to consider how this Type IIA vessel information relates to the other artefacts excavated at each site and to their stratigraphy. After all, the level of archaeological value afforded a contextless vessel is necessarily limited.

And what of the situations in which the information that you seek is not available on the internet? It is not uncommon for an archaeologist with an interest in a particular category of a project’s material assemblage to travel to the location of that project’s material record so that they may physically inspect it themselves. Often, this is because the material has never been studied. However, it can also be required because the material has been investigated but these observations have never been published. The empirical data created by archaeologists, and which plays such a fundamental role in the construction of their interpretations and narratives, tend to be overlooked by traditional forms of archaeological publication. In our example case, it might well transpire that Site A and Site B have been published as monographs or as a series of articles but it is certainly not a foregone conclusion that contextually-rich information concerning these Type IIA vessels will form part of these outputs.

The point of this thought experiment has been to highlight the levels of complexity and challenge involved in even the most superficially simple archaeological data gathering exercise. Any archaeologist reading this text will recognise the patterns of research that are described here. They will see – in these time-consuming and often frustrating stages – reflections of their own individual workflows. No doubt there will be differences as well but the essence of the method will be the same. And so it is around this subject of how archaeologists go about the fundamental task of gathering their data that this thesis is primarily concerned. My opening premise is that this series of approaches, which are so commonplace throughout all archaeological practice and which affect all stages of the process and all agents involved in that process, is fundamentally flawed.

The process is flawed for a number of reasons and it is the aim of this thesis to identify these reasons and to consider what might be done to address them. A central premise of this work accepts that archaeology is and indeed perhaps always should be a challenging and necessarily incomplete process. However, there is a substantive difference between the difficulty that serves to sharpen the blade of the researcher’s intellect and that, which
acts only as a retardant of epistemological innovation. There is no doubt that archaeology has largely welcomed the introduction of digital practices into its fold, particularly in the last decade. Indeed, it would be difficult to identify a single aspect of the archaeological method which has not been affected in some way by the use of digital methodologies. However, it is also true (and I believe reasonably uncontroversial) to say that the way that archaeology has used these new digital tools as platforms on which to represent, disseminate, collaborate over, and enrich the output of its processes has followed a distinctly disjointed trajectory. The once mooted standard for coordinating archaeological digital data information systems has now long since fallen by the wayside and to an extent the current hybridity of methodological and representational practice, which has resulted in the research difficulties illustrated in the opening example, is both product and testament to this fact. As such, the second working assumption of this thesis is that digital data heterogeneity is here to stay within archaeology and that as a result, any system that is designed to ease the workings of digital archaeological research must be cognizant and reflective of this fact.

On this basis, the primary solution proposed by this thesis to accommodate and indeed embrace this archaeological data heterogeneity is to promote the use of a data representation model known as the Semantic Web in as many facets of the archaeological method as is applicable. We will thoroughly examine what the Semantic Web means in the coming chapters but for now it is sufficient to say that it is a model for the publishing and consumption of digital data, which promotes its re-use and links to other data. The principle that archaeological research in its broadest sense has much to benefit from Semantic Web adoption is now relatively well established, with numerous scholars having explored aspects of this coupling (Byrne 2008; Isaksen 2011; Wright 2011). However, despite this academic interest and the fact that the principal archaeological data institutions are now actively promoting the use of Semantic Web approaches, the paradigm has yet to truly ignite the imagination of the working archaeologist and to reimagine the landscape of archaeological research practice. As such, while this thesis reviews the basic technical characteristics of the Semantic Web as they correspond to the archaeological need, an effort has been made throughout the project to dig beneath the surface layer of this functionalist viewpoint in order to assess the Semantic Web from a more consciously critical perspective.
Chapter 1 - The problem with Archaeological Information Systems

In that respect, a key difference between this thesis and related works is the focus placed on the archaeological data consumer in this knowledge environment. There has been a tendency by previous studies in this area to sideline the user perspective, in favour of presenting the Archaeological Semantic Web primarily as a concern of archaeological knowledge production (Isaksen 2011 p. 163). I argue that to neglect to consider the Archaeological Semantic Web consumer’s needs, expectations and experiences serves to perpetuate its indifferent reception within this community. To put it bluntly, it will make not one iota of a difference to the success of the Archaeological Semantic Web were millions of investment dollars, euros or renminbi to be ploughed into the building of new Archaeological Information System infrastructures, if consumers decide to reject the work practices imposed by these systems. Archaeological consumers need to be shown and be willing to buy into the benefits that come with Archaeological Semantic Web adoption and not be constantly warned about the responsibilities that it brings. This means that the field’s evangelists should start demonstrating how real world archaeological questions can be addressed using Archaeological Semantic Web techniques. The Archaeological Semantic Web has moved beyond the prototype phase and is now a normal Kuhnian science; this transition and its implications need to be communicated more effectively to the grassroots archaeological community. To develop this perspective, this thesis proposes a model of the Archaeological Semantic Web user, based on the findings of recent sociological research on the topic. It also reflects on how this model might fit within an existing understanding of digital epistemology and 21st century archaeological theory.

By working through these ‘bigger’ questions, this thesis has also had the opportunity to take in a number of related smaller studies, and each of these has contributed to the overall narrative. It asks the question whether it is possible for smaller, less well-funded archaeological projects to become contributors to the Archaeological Semantic Web, and the Priniatikos Pyrgos Project’s Archaeological Information System is presented in this context. It also investigates how the use of a Knowledge Representation system such as the Archaeological Semantic Web influences other aspects of the field’s research method.

1 Note that a convention of this thesis is to refer to the part of the Semantic Web that is related to archaeological and cultural heritage datasets as the Archaeological Semantic Web. This terminology is explained in Chapter 2.

2 This thesis can also be read as something of a reaction against the view that DH can only be understood as a large-scale project research environment. See Van Zundert (2012) for an argument against the value of pursuing ‘big’ DH research.
We look particularly at how the acts of finding and visualising data are two highly interrelated practices.

**Thesis structure**

To begin, Chapters 2 and 3 introduce the Archaeological Semantic Web from an historical, theoretical and functional point of view. In Chapters 4, 5 and 6 we consider how the Archaeological Semantic Web model might be implemented in practice. Chapter 4 describes the building of the linkedarc.net Archaeological Information System. Chapter 5 explains how the linkedarc.net system was populated with real archaeological content taken from the Priniatikos Pyrgos Project. Chapter 6 describes how this resource, as well as other Archaeological Semantic Web resources, might be consumed by an archaeological researcher. In the penultimate Chapter 7, we adopt a more consciously critical perspective on the model, asking whether an alternative system could not be used to garner a comparable set of results. Chapter 7 culminates by presenting the thesis’s central contribution; that is a sociologically and epistemologically informed model of Archaeological Semantic Web use. Chapter 8 in conclusion draws together the various narrative strands of the work, as well as speculating on the future paths of the Archaeological Semantic Web and of this particular research project.

**Digital practice within the humanities and archaeology**

Before we delve into the body of our enquiry, we need to carry out a number of housekeeping duties. The first and foremost of these tasks is to establish a context for this study. Chapters 2 and 3 will more specifically situate the Archaeological Semantic Web project within the context of developments in the internet and Knowledge Representation space. However, before attempting this, we must first address the subject in broader terms. Therefore, we begin with a discussion of the digital humanities, the academic field to which this research project belongs.

**What are the digital humanities?**

The simplicity of this question belies the nuance, complexity and range of possible answers that it provokes. On first impression, it would seem self-evident that the digital humanities or DH, as it is also known, is an interplay of two component parts: the humanities and the digital medium. Given that the latter of these is used in its adjectival form and, therefore, affects the nature of the former, it is perhaps reasonable to read the humanities element of the mixture as being the more important. But what aspects of
humanities does this refer to? Most humanists announce themselves in more specific terms than as simply ‘humanists’. They are English literature specialists, archaeologists, linguists, historians and so on. It is clear that the humanities are a broad church. Digital practice, as will be illustrated in this thesis, is also an extremely varied activity. It can denote word processing and email writing, video editing and writing code, blogging and much more besides. It certainly could not be claimed, however, that the mere undertaking of these tasks in themselves qualifies one for membership of the DH community (Unsworth 2013 p. 36). Most ‘real’ DH practitioners would baulk at such an idea, and it is certainly the case that not all humanists, who happen to write emails or edit the odd piece of video footage, would even want to be associated with this label.

One can also look at the digital in terms of the ‘objects’ that it produces (thereby considering these outputs as a form of materiality (Manoff 2006)) or as it manifests itself as a process (Coble et al. 2014). It is also possible to take the view that the digital, as a technology, can be determinative, that it can have a form of agency (Latour 2005 pp. 70–71; Rockwell 2010). It is evident, therefore, that delineating what is and what is not DH is no simple matter and, because of that fact, exercises in DH definition have become a sort of sub-genre of the field, replete with its own particular terminology and publications (Sample 2013 pp. 259–260). We will consider aspects of DH throughout this work, sometimes explicitly but more often than not these must be read between the lines of the discourse. Hopefully, the content and form of this thesis might also contribute something to the definition of what is still very much a juvenile member of the academy. Having said that, we first need to become better acquainted with the established views on the subject, so that we can enter the coming chapters with our eyes wide open to the ‘winks and nods’ of disciplinary inference. In that context, we will now take a necessarily speedy and regrettably limited tour through the development of DH as an academic discipline.

---

3 Another example of the popularity of these attempts at defining the field is the Day of Digital Humanities initiative. On a single day every year since 2009, Day of DH has been asking its participants to put into words their understanding of the field (Terras et al. 2013 pp. 2009–2012).

4 I use the term ‘juvenile’ in the best possible sense; that it is willing to take risks, to explore and to do work because it is interesting and not because it merely fits within a defined category of academic agency.
The three waves model

In the edited volume *Understanding Digital Humanities* (Berry 2012), the various attempts to define DH are framed within a consensus view that the field has gone through two major phases or waves of development. The objective of the volume is to chart a course into a new epistemological framework for DH, which they call the ‘computational turn’ (Berry 2012 pp. 4, 14). The three wave structure is not the only paradigm that has been proposed to explain the development of the field – some argue for instance, that the original tenets of DH still remain largely sound (Drucker 2009) – but it is the one that we will follow here, principally because it conveniently encompasses a wide range of views put forth concerning the form of DH to date.

Broadly speaking, the first wave of DH was focused on the ‘building of infrastructures in the studying of humanities texts through digital repositories, text markup and so forth’ (Berry 2012 p. 4). Following Berry, early definitions of DH tended to emphasise the primacy of text as an object of study within the field (Frabetti 2012 p. 162; Svensson 2009). This text-focus was either the cause or the result of the close association that DH studies enjoyed with modern language literature and linguistics departments in its early years. The pioneers of the field used computers to perform computational tasks that would have otherwise taken a human an impractically long period of time to complete.

Essentially, almost all DH research conducted up to end of the last century was concerned with the finding of patterns in data, which was predominantly in the form of text (Dixon 2012). This does not mean to suggest that the first wave of DH was atheoretical – see Unsworth’s attempt to define the field as a type of ontological

---

5 The phrase ‘computational turn’ was the title of a workshop held in the Swansea University in 2010. The intention of the workshop was to speculate on the future direction of field. Presumably the phrase is a reference to the so-called ‘linguistic turn’, which came to identify the language-based focus of the philosophical movement of the 20th century. See Chapter 3 for more on this subject.

6 This textual focus needs to also be understood in the context of the history of humanities research in more general terms. While non-textual subject media are not unknown to the humanities – take the activities of art history or media studies as examples – it is, I believe, uncontroversial to state that the vast majority of humanities scholars have since the creation of the modern university used text as their primary information medium.

7 It is generally acknowledged that the DH tradition first begins with the pioneering work of the Italian Jesuit priest, Father Roberto Busa, who in 1949 paid a visit to the CEO of IBM, Thomas Watson, with the idea that Mr Watson’s machines could be used to help Busa with a project (Hockey 2004). Busa wanted to compile an index of the 11 million Medieval Latin words encountered in the complete works of St Thomas Aquinas. This formidable task eventually resulted in the publication of a series of volumes many years later and only in 1992 did Busa disseminate the work in digital form using the CD-ROM media format.
commitment, as an example (Unsworth 2013) – but rather that the digital or computing component of the method was mostly assumed to be epistemologically neutral or irrelevant.

During this first period, DH was not yet known as ‘digital humanities’. Its original preferred label, ‘humanities computing’ had placed its emphasis on the computing component of the relationship (Svensson 2009). While one should not fall into the trap of reading too much into the meaning of labels, the fact that both words in the humanities computing term are nouns is perhaps informative. It suggests that humanities computing placed the humanities and computing within two separate spheres of activity, and as it happened that is the way that much of the work in the space was carried out. It was not a collaborative relationship in the sense that we understand that term today. In other words, this was not a relationship of equals. To illustrate this point, Andrew Prescott recounts in his blog how he was struck during the first months of his tenure in the Department of Digital Humanities at King’s College London ‘by how often [his] external academic partners assume that they are the driving force in the collaboration’ (Prescott 2012). In this first iteration of DH, the humanists ruled the roost. Their instructions were interpreted and achieved by what they viewed as their computing technicians. As an outcome of this attitude, the idea of the digital humanist as a ‘tool’ of the humanities intellectual came into being (Rockwell 2012 p. 252).

The term ‘digital humanities’ did not come into common usage until relatively recently. Hayles ventures, on the back of Ramsey’s account, that it was Drucker, Unsworth and McGann who first started using the term during the late 1990s to describe the work that they were doing at the University of Virginia’s Institute for Advanced Technology in the Humanities (Hayles 2012 p. 43). The new label was an attempt to correct the dichotomous and perhaps even antagonistic relationship that they perceived to exist between the humanities and computing communities at that time. The second wave of DH did not begin with this rebranding but the seeds of its theoretical structure are found in this event. Second wave DH came about as a result of changing technology. The arrival of Web 2.0 brought with it new forms of digital content (blogs, micro-blogs, video blogs, wikis and so on),8 which very quickly began to appear as the subjects of DH

---

8 See the discussion of Web 2.0 in Chapter 2 for more on this technological and sociological shift.
research. As such, second wave DH is distinct from its predecessor primarily in its choice of subject matter. ‘Born-digital’ artefacts came to be viewed as suitable for academic study, whereas first wave DH was solely interested in the digital correlates of the ‘real’. The types of practice used in the second wave had also evolved. Hayles (2012 p. 43) describes the second wave methodologies as being ‘multimodal’, in the sense that DH practitioners were now borrowing heavily from other fields, most notably media studies and the nascent digital media studies. Evans and Rees see the change from first to second wave as being primarily about a shift of methodological interest from quantitative to qualitative. For them, the ‘digital [in second wave DH] is embedded within our very perceptions of the humanities’ (L. Evans & Rees 2012 p. 23).

At a disciplinary level, second wave DH promotes communication across traditional academic boundaries. Academics first referred to this type of practice as interdisciplinarity but now even this is deemed to be too conventional a term for what DH is trying to achieve in the academy (Liu 2014). Berry calls for a DH that is trans-disciplinary; that is an end to the clearly delineated boundaries that define our current academic structures (Berry 2012 p. 13).

A significant outcome of this move towards trans-disciplinarity has been an imperative to introduce new management structures into the practice of DH (Edmond 2016). It is clear that there are significant differences to be found between the traditional humanities ‘lone research scholar’ model and the ‘multi-agent team’ structures that form the nucleus of much recent DH research. Large teams need to be managed effectively if their combined efforts are to be rendered greater than the sum of their parts. DH projects also have to contend with the management of budgets and deadlines. Large DH project dissemination strategies are also likely to be more complex than the approaches to publication that have defined the traditional humanities research model for many centuries. Because of these new demands DH projects now require the involvement of a new breed of academic who must be skilled not only in the facilitation of a DH team’s collaboration but in other less-visibly academic duties as well. These DH project managers must be able to administer project finances and to oversee project plans. While other project members are able to concentrate on their own specific tasks, the DH project manager must be constantly cognizant of the bigger picture, of how all of the individual components fit together into a coherent whole. An upshot of this is that they
need to be constantly aware of the wants and needs of their consumers or users (Warwick et al. 2007, 2009, 2012). As a result, DH project managers now quite rightly try to engage with their users at all stages of the project’s development. When added together, all of these new demands mean that the large-scale second wave DH project manager must have many of the skills that are more in keeping with business practice than have been traditionally associated with humanities research models to date.

In 2009, Schnapp and Presner (2009) produced a manifesto that set down the path they believed second wave DH should follow. They proposed that the field adhere to five basic tenets: that it be qualitative, interpretive, experimental, emotive and generative. While adventurous in its outlook, it also must be said that the Schnapp and Presner vision for DH is also largely consistent with the basic principles of postmodernist thinking and, more specifically, with critical theory. Each of the manifesto tenets is essentially an argument for the value of hermeneutics. As such, second wave DH is about employing ‘heuristic digital methods’ (Rieder & Röhle 2012 p. 69) to help draw out new interpretations from traditional and new media, or as Liu (2008) puts it, digital methods become ‘tools that we think through’. Second wave DH is anti-positivist and supportive of a constructivist view of knowledge creation (L. Evans & Rees 2012 pp. 28–30).

And last, there is third wave DH, which, as we have already observed, is also referred to as DH’s ‘computational turn’. Depending on your point of view, DH is either entering into this phase or it is an aspirational paradigm, which may never truly come to pass. In Berry’s mind, the third wave is about essentially problematising the method of DH (Berry 2012 p. 5). He suggests that we need to ‘look at the digital [sic] component of the digital humanities in the light of its medium specificity, as a way of thinking about how medial changes produce epistemic changes’ (Berry 2012 p. 4). This focus on the medium as being as, or perhaps even more, important than its content is an idea that first took root in the writings of the media scholar, Marshall McLuhan in the 1960s and 1970s (McLuhan & Fiore 1967). It builds upon the basic assumption that every aspect of an epistemological system, be it the content, the message or the messenger, needs to be critically challenged. Accepting this model has a number of implications, which are relevant to DH. It is in conflict with an objectivist view of the world. This is an epistemology that has often been implicitly aligned with the computer, in the sense that
these machines are ‘computers’ of objective data and appliers of rational axioms (Rieder & Röhle 2012 p. 72). McLuhan’s thesis also raises deep concerns about the rise of the visual medium in the construction and dissemination of 21st century knowledge, which have since been developed by other scholars (Berger 1972; Sontag 1977).

Scattered throughout these three epochs of DH practice are a number of recurring themes, which we have yet to reference. We did not mention, for instance, the variable of scale, which crops up in a number of different DH contexts. Take Moretti’s ‘distant reading’ paradigm, in which he argues that by using digital technology it is now possible for a single scholar to ‘read’ the contents of more books than was ever previously possible (Moretti 2013). Scale is also fundamental in the analysis of social media data. DH practitioners can now quite easily analyse millions or even billions of units of information sourced from these channels (Manovich 2012). When dealing with this amount of data, comparisons with pre-digital research are almost pointless; it is to compare apples and oranges. Another habitual debate within DH surrounds the question of computer code. Should a DH practitioner be able to write code? Ramsey (2013) thinks that they should be able to but others disagree (Sample 2013). And if they cannot, should they at least be able to read code (Rieder & Röhle 2012 p. 77; Sample & Vee 2012)? And in that sense, might it be worthwhile to consider code as an object of study in itself, similar to the text of a work of literature (Ramsey & Rockwell 2012)? Code, as a process of making, raises the related question of DH and building (Hayles 2012 p. 60; Liu 2014). We will develop this understanding of DH further when we propose a model for the digital humanist later in this chapter.

The membership and audience of DH are also issues that have generated interesting debates over the years (Edwards 2012; Rockwell 2011; Warwick 2012). Should the community work towards an inclusive (Terras 2013 p. 266) or an exclusive membership policy (Rockwell 2011)? How do DH audiences differ from those of the traditional humanities? Do different audiences necessitate the use of different language registers? The subject of DH output is also topical. Should DH publications be made to fit within existing dissemination norms (Terras 2012 p. 180)? We have already mentioned the importance that is now placed on the collaborative effort in producing DH knowledge. How is this type of work to be assessed and indeed rewarded within a system that
currently prioritises the solitary scholar and the traditional publication mechanism (P. Cohen 2010; Spiro 2012 p. 23)?

These are a sample of the ‘meta’ questions that currently occupy the minds of DH scholars. Unfortunately, we do not have the luxury in this text to do more than to briefly reference them but this exercise should at least serve to intellectually situate the current work and to highlight how many of these matters still remain contentious to practitioners of DH. This is important, as this thesis is as much a product of DH, as it is an archaeological treatise. And this model of stepping from one academic frame of reference into the next is another key facet of the DH paradigm. Even though the subject material for this thesis comes almost exclusively from an archaeological domain, the approaches used to observe, model and interpret this data are also witnessed in the workflows adopted by historians, linguists, geographers and others. By allowing (or forcing) digital humanists to move outside of their disciplinary comfort zones, it exposes them to new ways of thinking (not always digital), which will ideally come to affect their practice and ultimately their parent fields.

A model of the digital humanist

The question of how best to define a member of the DH community is, I believe, key to gaining an understanding of the DH space. Generally speaking, the inclusive model favoured by Terras would seem to have more to offer than its more exclusive alternative. Having said that, a liberal entry requirement needs to be understood in the context of the field’s subject matters and not in the types of method that are applied to these materials. In order for the field to maintain some degree of intellectual coherence, there needs to be a degree of prescription imposed on the types of method that its practitioners employ. This does not need to, and indeed should not, fall into the trap of methodological prescription, as to be overly rigid in this regard would stymy the field’s development. But there are certain characteristics that do serve to differentiate the DH scholar from, for instance, the traditional humanities researcher who still inhabits the same digitally enabled world as the digital humanist.

In this section, we outline a DH practitioner model that adheres to these basic principles. This model should be understood as a foundation on which the remainder of this text can be read. It is best to view this model as something of a folksonomy of DH practice as I have experienced it over the last number of years. As with any model, it would be a
mistake to view the DHer model (as it has been named) as a universal objective reality for it is clearly not an exhaustive list of all of the possible ways that DH practice can be carried out. If the preceding history of the field has shown us anything, it is that DH is a decidedly difficult discipline to delineate. Often when one inspects a particular project or other, it may not be entirely apparent why the work qualifies for inclusion within the sphere of DH research even though this qualification is never in question. In these cases, it is perhaps the accumulation of a multitude of inferences that serves to satisfy the entry criteria, with no one feature being deemed sufficient to bring the others across the DH line. Bearing this all in mind and in the knowledge that the following taxonomy will no doubt invite as many objections as it does endorsements, let us proceed.

**DHer 1 – the baseline digital humanist**

The first type of digital humanist (or DHer 1, as we will refer to him or her) is the baseline practitioner of DH. The types of research question that interest DHer 1 might be in line with questions asked by the traditional humanities researcher or they might address subjects that are related to the digital realm itself. However, what clearly marks them out as distinct from their traditional colleagues is their use of a broad range of digital tools and methods. These form the basis and core of their research praxis and are in that sense indispensable. For instance, DHer 1 will have a generic level of knowledge in the use of techniques such as the storing of data in relational databases and, as such, they will know how to write SQL queries that can be used to intelligently access this information. They will be skilled in the use of digital research methodologies and will also be comfortable with distributing their research using digital dissemination tools (social media channels, digital journals, et cetera). They may also possess a basic level of competence in writing computer code. For instance, they may be able to build simple client-side web pages using JavaScript and Open-source Software such as jQuery and Leaflet. DHer 1 may or may not form part of a collaborative research team.

**DHer 2 – the tool specialist**

DHer 2 is similarly focused on the use of digital methods but their level of engagement with specific digital tools is much more pronounced. For this reason, DHer 2 is the advanced tool user of the group. DHer 2 will be highly skilled in the use of a single digital tool or method. Typical examples of these methods are Text Encoding Initiative,

SQL stands for Structured Query Language. It is a formal language used to interface with a database. SQL is discussed in greater detail in Chapter 7.
Geographical Information Systems and 3D data capture and representation. The distinction between DHer 1 and DHer 2 is a matter of training and the manner in which they employ this knowledge. DHer 2 has invested a substantial amount of time and effort in learning the esoteric ways of mastering their chosen tool. As a result of going down this path, DHer 2 will spend the majority of their research time using these tools for a variety of different purposes. Because of this tool focus, DHer 2 is likely to be more flexible in terms of the subject matters that they engage with. For instance, one month they might employ their skills to address an English literature subject matter, while in the next their attention is turned to a historical topic. DHer 2 is likely to work as part of a collaborative team. As such, their engagement with the research question to hand will be heavily mediated by the contributions of other members of the team, who may have a higher level of domain knowledge.

**DHer 3 – the builder**

DHer 3 is a creator of DH products. They have advanced knowledge in the writing of code and in the building of sophisticated digital infrastructures. DHer 3 is more likely to work as part of a multidisciplinary team but it is also possible that they may work alone. What distinguishes DHer 3 from a computer science engineer is their parallel knowledge of the humanities. For this reason, it is likely that some sizeable proportion of their education has been conducted in a school of humanities. Members of the DHer 3 group build tools that will be used by humanists and other digital humanists. The reason that the DHer 3 category is needed within the DH community is that their existence goes against the unhelpful ‘computer science outsourcing’ model, which typified the early stages of DH. Humanists have an understanding of humanities subject matters that would not necessarily be the case were an engineer to be used to develop these digital systems. They speak the language of humanities and so are more likely to successfully transform the conceptual designs of humanist colleagues into realised digital forms. The ‘DH practitioner as maker’ debate has been referred to already but it needs to be emphasised that this is a contentious area of discussion within the field. Currently, the reality is that the majority of DH practitioners do not code but in time we should see the amount of digital humanists who are able to code increase. It is likely that the DH space of 20 years time will include a much higher proportion of DHer 3 practitioners than is currently the case (Hayles 2012 p. 59).
Chapter 1 - The problem with Archaeological Information Systems

**DHer 4 – the project manager**

The job of the DH project manager is comparable to that of the manager of any complex project, academic or otherwise. In fact, one of the fundamental characteristics of the DH project that marks it out from other humanities research projects is that the former is far more likely to transcend societal boundaries; for example from academia into the spheres of industry and the public sector (Burdick et al. 2012 p. 48). This is a positive development in many ways but it also places new demands on the DH project and the skillset of DHer 4 reflects this more complex working structure.

DHer 4 needs to be adept at the creation and execution of a project plan, which may in some cases run for many years and involve multiple phases. They must also be expert people managers. DH projects now typically engage with different types of actors (the general public, policy makers both at home and abroad, academics, and so on) who have diverse sets of needs (Burdick et al. 2012 p. 112). DHer 4 must be able to create an environment, which allows scholars of often quite different disciplinary backgrounds to work effectively together (Siemens 2009 p. 226). Communication is, therefore, a vital cog in the running of any successful DH project. DHer 4 must enable communication to flow between these different domains, which more often than not will use their own domain-specific languages. They must also act as the spokesperson for the group, as the interface between the internal workings of the project and the outside world (Cetina 1999 p. 180).

There is far less onus on DHer 4 to have the sort of specialised digital tool knowledge that is characteristic of the previous three digital humanist types. In the same way, specific project domain knowledge is not an absolute necessity for DHer 4, although, in the context of their role as project interface, this requirement becomes much more relevant.

**DHer 5 – the domain knowledge specialist**

The final digital humanist type included in our model is the domain knowledge specialist. Superficially, DHer 5 appears to be the same as the traditional humanities scholar. They have acquired a high level of knowledge about a specific category of humanities content through the standard (graduate and postgraduate) academic channels. However, what marks them out as special is their ability to work within the interdisciplinary teams of larger DH projects. They will, therefore, thrive in a collaborative environment, enjoying
the opportunities that come with working with designers, computer programmers, policy
makers, legal experts and other domain knowledge specialists. While they themselves are
not required to be skilled in the use of digital tools and methods, they will nonetheless be
able to communicate with those team members who do have these skills. This distinction
follows Rieder and Röhle’s (2012 p. 77) argument that not all digital humanists need to
be able to write code but they should all at least be able to make some sense of it and to
engage in dialogue around it.

**Digital archaeology and Archaeological Information Systems**

While techniques favoured and developed within DH are also being applied in
archaeological contexts, it would be inaccurate to say that the history of the use of
computational methods by archaeologists is subsumed into the history of DH. While
both now share much common ground in terms of the personnel involved, the technical
strategies followed and in the theoretical structures that underpin these approaches, their
histories have developed along separate paths and are in fact quite distinct, as we will
now attempt to elucidate.

The history of digital archaeology\(^1\) begins in the late 1960s when archaeologists first
started to use computers as datastores and as mechanisms for conducting statistical
analyses that were up until that point impossible to achieve using human computation
alone (Eiteljorg 2004). These early experiments were rare for their time and would no
doubt have been viewed with suspicion and their motivations and logic questioned by
the peers of those conducting them. Despite the vast costs and the great tedium involved
(Daly & Evans 2006 p. 6), interesting results did start to appear, most notably in the
statistical work of archaeologists such as Deetz (1977) and Longacre (1970). It was also
during the 1970s that the first conferences on the subject of digital archaeology were held
(Eiteljorg 2004 p. 22). The inaugural Computer Applications and Quantitative Methods

---

\(^1\) The use of computational methodologies within archaeology has not only been referred to as
‘digital archaeology’. ‘Computational archaeology’, ‘archaeological informatics’,
‘archaeoinformatics’ and latterly ‘cyber-archaeology’ (Zubrow 2010) are terms that have been
used extensively within the community and continue to be used to describe this general category
of practice. A Google search for ‘digital archaeology’ shows that it is also used to define the
activity of investigating and preserving the early digital record of the World Wide Web (Boulton
2014). And so, while I am fully aware that there are epistemological implications in using each of
these terms in the context of archaeological research, we will set these matters aside for now and
for the sake of simplicity, use ‘digital archaeology’ to refer to all digital practices conducted by
archaeologists, computer scientists or any other category of researcher as they apply to matters of
archaeological relevance.
in Archaeology gathering took place in the University of Birmingham, England in 1973. CAA would go on to perform a unique function for the field from that point on. However, despite these advances, it would take a number of years and, specifically, a number of key technological developments before the use of computer technology became truly democratized within archaeology. The most significant of these technological catalysts was the move away from mainframe computing towards the personal computer, which started during the late 1970s and accelerated in the 1980s (Berndt & Rappaport 2001 p. 268). The PC was cheaper\(^\text{11}\) and more portable, and Zubrow characterizes the impact of this technological shift on the social reality of digital archaeology as being an increase in the “individualism” in machines’ (Zubrow 2006 p. 16).

Since that change, interest in using digital archaeological techniques has grown from strength to strength. In retrospect, it is possible to categorize these types of project into five basic groups: recording and representation of archaeological information, statistical analysis, data modelling, virtual reality and mixed reality applications, and finally, digital dissemination (Daly & Evans 2006 p. 5). Digital archaeology is now also fully integrated into the didactic fibre of the field. While certain ‘hubs’ of digital archaeology exist (for example, at the University of Southampton in the UK and at Stanford University in the US), most, if not all, archaeology departments now run modules, which address aspects of digital practice. Consequently, these methods have become ubiquitous elements of field and lab archaeology. For instance, one of the absolute prerequisites of any archaeological field project is the inclusion of a Geographical Information Systems specialist in its team, and we see the development of an analogous canon of digital skills (3D modelling, digital photograph editing and so on) being introduced into post-excavation practice as well. Whether considered within its commercial or research guises, the application of digital techniques within archaeology has reached the point of normalization.

\(^{11}\) The price of the earliest personal computer models dwarfs its present day equivalent. If we compare two like-for-like desktop personal computer models, with the first being purchased in 1983 and the second in 1999, the former is 384 times more expensive in real terms than the latter (Berndt & Rappaport 2001 p. 271).
Archaeological Information Systems\textsuperscript{12} represent one combination of digital activity and artefact, which is firmly associated with the rise of digital archaeology, and they are also the primary focus of this thesis. Usually defined within the context of digital recording practice, the Archaeological Information System’s role has become broader as its supporting technologies have evolved and the imaginations of its creators have expanded to exploit these new technological possibilities. Having said that, the primary task of the Archaeological Information System has remained to organise and store the empirical observations made by archaeologists during excavation and post-excavation and the interpretations that are ultimately derived from these data. Of course, this does not mean to say that an Archaeological Information System must be delivered as a digital solution. In fact, we will show in this thesis that it is far more common for archaeological projects to use paper-based Archaeological Information Systems, at least at the beginning of their life. Regardless of its medial form, so fundamental is the Archaeological Information System to the workings of an archaeological project that to imagine one without the other is perhaps impossible.

The first digital Archaeological Information Systems were created in the 1970s, as primitive (from today’s standards) database applications. These were originally designed to run on mainframe computers but as we have seen it was soon necessary to port their functionality over for use on the more flexible architecture of the personal computer (Eiteljorg 2004). While the technology of the Archaeological Information System is clearly of importance in the context of the services that it can deliver, it is just as important to consider the Archaeological Information System from a more holistic perspective. The way that archaeological information is digitally formalised is governed by the laws and practices of a branch of information science known as Knowledge Representation (Sowa 1999), and while KR is often presented from a narrowly functionalist point of view, this thesis bucks the trend by addressing its wider causes and effects. As such, Chapters 3 and 7 present the KR decision-making process in respect to its associated ontological, epistemological and sociological contexts.

\textsuperscript{12} Alternatives to the ‘Archaeological Information System’ term occasionally appear but this would seem to be caused more by a lack of awareness by the authors about the existing scholarship on the subject than by any significant change of sense intended. As such, I have chosen to follow the example set by Carver (2005) and others (Joseph et al. 2004) and to exclusively use Archaeological Information System for the remainder of this thesis.
One aspect of the Archaeological Information System, which invites as much functionalist debate as it does questions of a more theoretical nature, is data standardisation and this is a topic that will play a substantial role in the narrative of this thesis. Archaeological Information System standardisation defines the principle that the degree to which an individual project can decide how its data is structured is limited by external dictates. So, for instance, imagine that an archaeological project wishes to define a particular category of information that will record stratigraphic information for their project. If the project chooses (or is compelled) to adhere strictly to an external Archaeological Information System standard, then the form of this stratigraphic record will be decided not by the project’s members but by some external body. Advocates of the standard would argue that this system allows for researchers to consider the contents of more than one archaeological project at a time. This increases its value at a pragmatic level but critics would contend that the functional gains are offset and overshadowed by the loss of individual project freedom, which has implications for the project’s epistemology. The arguments are more nuanced in reality than this and most Archaeological Information Systems will impose some level of standardisation of practice onto its community of users. The debate centres on the degree of adherence and it is a matter that will structure the form of many of the arguments developed in the coming chapters.

As was mentioned at the beginning of this section, digital archaeology and DH have yet to develop the close working relationship that would reflect the theoretical and practical commonalities that link the two fields (Watrall 2012). There are a number of possible reasons why such a developmental distinction might have come to pass. It is possible, for instance, that there is a sense within the two communities that they are both competing for the same limited funding and other intuitional resources. It also might be the case that DH’s historical focus on mainly textual content is at odds with archaeology’s interest in a much broader range of material culture (Dunn 2011 p. 95). I would also argue that the large-scale trans-disciplinary team model that has become popular in DH is not yet the norm (and it may never become the norm) in digital archaeological circles. And yet despite these differences, in the light of Terras’s call for a DH that espouses a ‘Big Tent’ model of inclusivity, it is not hard to conclude that both communities are currently missing an opportunity when it comes to learning from the experiences of the other. DH and digital archaeology have much to gain from the establishment of a more
collaborative working relationship. Personally speaking, I identify as a digital humanist but at the same time, this statement should not neutralise the fact that I also work within the archaeological community as an excavator, researcher and consumer of materials derived from archaeological practice. And in the same measure, it should not, and I would hope does not, preclude my involvement in the activities and debates of the DH community.

**Archaeological data conceptualisation and data terminology**

A key topic that needs to be addressed before going any further is the subject of archaeological data conceptualisation and how we use language to define the concepts and relationships that constitute archaeological knowledge. After all, this is a thesis that is essentially concerned with the flow of data through the epistemological channels of the archaeological discipline, and so the question of ‘what is archaeological data?’ must be addressed. My own position is that there can be no one universal answer to this question; how you come to conceptualise archaeological data and all of the metaphysical structures and more tangible outputs that grow out of that understanding are dependent on your own particular world view, which is a product of your nature, your culture and your experience of other natures and other cultures. In that light, it is important that I, as the author of this work, now set out my stall and define exactly what it is that I mean by the term, archaeological data.

The philosophy of archaeology has for the last number of decades been typically (and problematically) divided into two camps: processualism and post-processualism. We will present a more nuanced history of this dynamic in Chapter 7 but for now we can say that the first largely equates to the mores of positivism and the latter to those of post-modernism. Positivism essentially posits that truth is obtainable and the various factions of post-modernism agree that it is not. For the archaeological processualist, this implies that the past is accessible through the medium of the material remains of past agency. For example, were a processualist archaeologist to uncover a fine bone comb during the excavation of a Viking settlement, this discovery could be used to open up a portal into a series of discreet past realities surrounding the biography of the comb. A post-processualist would be much more circumspect in terms of what they felt they could say about the people and societies who had once interacted with the artefact. They would be constantly aware that whatever theory was induced from the comb data was bound to say more about themselves and their own societies than that of any other.
Clearly, this divergence of views imposes two very different understandings onto the nature and value of data within archaeology. For the processualist, data can be trusted and used confidently as a source material upon which theories can be constructed. For the post-processualist, data is to be treated with caution, perhaps even mistrust. These views become even more entrenched when we introduce digital data into the equation. Processualists will see digital data as being reflective of the physical data, which in turn is reflective of past agency, as we have just seen. For the post-processualist, however, digital data introduces a new level of obfuscation, which serves to further distance the archaeologist from the past and, if anything, it re-enforces the subjectivity of the archaeological act.

How might the physical object sit within these two very different data models? Archaeology has traditionally been defined by the value that it places in the study of the object in its physical sense (Knappett 2014 p. 4701). It was long considered distinct from history because the latter focussed its attentions solely on the textual source, while archaeology was drawn to the study of the material record (Alberione dos Reis 2006). While this distinction is no longer considered as helpful as it once was (Bevir 2000; Isayev 2006), the gravitation of the archaeologist towards the idea of the object cannot be ignored and indeed it needs to be taken into account when we consider the world of digital knowledge. Throughout the coming chapters there will be mention made of digital entities, which would appear to have much in common with the archaeologist’s understanding of the physical object (Morgan 2012 p. 21). How then do these two concepts – the physical and the digital object – relate? Is the latter and more recent of these ideas built upon the other and, therefore, has the long tradition of thought that has surrounded the physical object come to influence our understanding of the latter?

My own opinion is that the answer to both of these questions is a resounding yes. I think that this understanding has come about in a natural, unnoticed and all-pervasive manner that is the hallmark of most human enterprise, which happens through a process of evolution and not revolution. I naturally think of a digital representation of a dagger that has been excavated from a site as being analogous to the physical object of that dagger. I
Imagine it as being related to other digital objects – such as the digital representation of its place of discovery, the date of this event and the archaeologist who performed the discovery act – but I also see the object as having characteristics (or metadata) – digital representations of its width, height, weight, colour, et cetera – that are enclosed within the confines of the dagger’s digital object. These are perhaps what one might term ‘natural’ views of the digital dagger object but while I hold these views I am also capable of appreciating that there are many alternative ways to conceptualise a digital object. For example, I also know that the digital dagger exists as a series of electrical charges stored on an integrated circuit within a computer’s memory. When viewed from this perspective, the idea that the digital record of the dagger’s characteristics are in some sense separated from the digital record representing the archaeological site, the date digital object or the archaeologist digital object no longer holds up to scrutiny.

Our understanding of digital archaeological data and the digital archaeological object is very much related to terminology. We all use words such as data, information, fact, truth, knowledge and interpretation frequently in our daily lives and in our academic discourses but it is not always clear how one term relates to the next and what, if any, the significance of choosing to use one over another has. As such, it is worth outlining what I understand by these data-knowledge terms (as I will refer to the group as a whole) and this should help in the reading of the coming chapters.

To begin, we must first consider the term ‘data’ itself. Kitchin tells us that data derives from the Latin verb ‘dare’, which means to give (Kitchin 2014 p. 2). As such, this implies that data involves the movement or communication of some thing from a source to a destination. Depending on the context, these sources and destinations will be different but one of the most common uses of data is in the context of the observation of phenomena. In this light, data can be understood as phenomena that are in some way captured or frozen in time and space. As we have seen, if you are a post-processualist, the model of data as a reproduction of a source phenomenon no longer stands up to

13 ‘Imagine’ is an important term here. The need to create visual images of digital objects in order to consider them is strong and yet it is a framework that the subject is rarely conscious of.
14 Compare this to Shanks’s (1997) definition of the term ‘naturalism’ as it is used within the field of archaeological photography.
15 The intention of adopting this personalised approach is not to present these definitions as universally acknowledged facts but rather to provide a guide with which the reader can navigate the coming text.
hermeneutical inspection and so we need to think about data as being entities produced as a function of environmental and subjective contexts (Gitelman 2013 p. 2). And in that sense, the models of data and representation begin to merge.

The data term is closely tied up with the concepts of facts and evidence and this network of relationships can be difficult to disentangle. Again, the emphasis here needs to be placed on the context of the use of these words. For instance, Rosenberg’s statement that ‘facts are ontological, evidence is epistemological [and] data is rhetorical’ (Rosenberg 2013 p. 18) defines the semantics of these terms as a function of the systems in which they operate. While one can argue one way or another as to the rights and wrongs of Rosenberg’s model, the important point is that she sets out explicitly her understanding of these terms from the outset. In this thesis, we will avoid the use of the terms ‘facts’ and ‘evidence’.

The term ‘raw data’ is, however, used frequently in this text and in other texts, which deal with KR topics. In general, the intention of most authors is to portray raw data as in some way removed from systems of interpretation. In this view, raw data is an objective fact that is the same for all observers. This is not, however, the view taken by this thesis and while I would accept that raw data is at the bottom of the data-knowledge ‘food chain’, I do not see it as being analogous to objective fact. Again, it is in my opinion necessary to consider raw data as a product of context and experience and, while this introduces challenges into the practical implementation of raw data in Archaeological Information System configurations, these need not be insurmountable, as is discussed in Chapter 7.

It is also possible to categorise data based on its formal and source characteristics. Does it refer to images or to quantifiable or more descriptive observations? Is the data being produced continually (i.e. from a real-time source (Berry 2012 p. 15), such as in the case of weather data) or does it represent an observational period that has now passed (for example, the data derived from the archaeological excavation of a site that is now no longer in existence)? Or, perhaps the data has no ‘real world’ referent as is the case of ‘born digital’ data (Rieder & Röhle 2012 p. 67). It is also possible to group data based on its relationship to a research infrastructure – for example, legacy data is different to data produced by contemporary research (Terras 2012 p. 177).
The type of essence that a unit of data wishes to record also has direct implications for its representation in a digital system. For instance, data can be serialised (committed to digital storage) as text, as a number, as an element in a set of values or as a combination of all of the above (Kitchin 2014 pp. 4–5). Chapter 5 considers the advantages and disadvantages of storing data in each of these forms but in a general sense, it needs to be understood that the form in which a piece of data is digitally stored has a huge impact on the types of uses that it can be ultimately put to. Generally speaking, the ease with which a machine can make use of a data type relates to its level of structure. More structured data can be more easily interrogated, whereas less structured data presents more of a challenge to machines. And as always, the implications of using one data form over another are not merely functional.

Despite the debate over the subtle differences that exist between the ‘lower-level’ units of data-knowledge terminology, the more abstract concepts of information and knowledge enjoy a broader level of consensus, with information typically seen as the next step up from data. Information is generally understood as the use or ‘processing’ of the data layer into something that is useful (Ackoff 1989). And knowledge is then a combination of various pieces of information. For now, we will say no more on these more distilled forms of data as Chapter 7 considers these matters in greater detail.

**My ideal reader**

Following the example set by McCarty (McCarty 2005 pp. 7–9), I thought that it would be a worthwhile exercise to attempt to define some characteristics of the ideal reader of this thesis. My intention here is not at all to scare off those readers who might not entirely fit this description but merely to make clear some of the assumptions that have come to structure this work’s writing. The first point to make is that this work straddles a theoretical and methodological space found across archaeology, computer science and digital humanities. There are chapters, which are entirely technical in tone and in content. These are necessary because the topics that we deal with – information systems, data querying, linking, sharing and visualisation – need to be understood first as technical entities. This also reflects the fact that the vast majority of the work that preceded the writing of this thesis was concerned with the overcoming of technical challenges. However, I also wish to avoid an outcome where the technical becomes an overpowering influence on the overall flow of the thesis’s narrative. To achieve this sense of balance,
other chapters generally eschew this technical focus and instead concentrate their attentions on presenting the Archaeological Semantic Web from a more humanistic and philosophical perspective. I am aware that this approach runs the risk of producing a work that is a ‘jack of all trades and a master of none’ but I believe that this is the risk that must be taken if a new narrative of the Archaeological Semantic Web is to be constructed.

The implications of this particular approach for you, the reader, is that you will be asked to shift from one mind-set to the next, possibly within the space of a few paragraphs. When appropriate, relevant technical texts are reproduced in the body of the document. These vary in type but mostly they relate to the matter of creating and querying Linked Open Data content. I have generally resisted the temptation of including any of the nearly 400,000 lines of code\textsuperscript{16} that have gone into the building of the linkedarc.net system and the related tools that have been developed throughout this project. Instead, the abstracted sense of these code modules is included in the body of the text as they are required.

**Thesis conventions and online resources**

The text that you hold in your hands is one half of this research project’s output. It is partnered with the linkedarc.net system, which is the digital realisation of the ideas presented in these pages. The linkedarc.net resource can be accessed using your web browser at [http://linkedarc.net](http://linkedarc.net) and it should be consumed in tandem with the text as you move from one topic to the next. Currently, linkedarc.net hosts the data of a single project: the Priniatikos Pyrgos Project.\textsuperscript{17} The data is made available as Open Data under a liberal re-use license and you are encouraged to engage with and use it as you wish. As well as providing access to the data of the project, the linkedarc.net web app also contains a number of additional digital resources. These delve further into aspects of the system’s design than was possible within the text. For instance, you can read supplementary material on the subject of implementing the CIDOC CRM data model

\textsuperscript{16} As of 3 September 2015, the total line count in all of the project’s code modules was 372,453. 20,459 of these lines were written for the linkedarc.net server, 179,324 for the linkedarc.net web app, 163,147 for the Priniatikos Pyrgos Textiles web app and 9,523 for the RDF Data Utilities web app.

\textsuperscript{17} The relationship between the linkedarc.net web app and server will be explained in detail in Chapters 4 and 5.
Wherever possible and appropriate, references to online resources have been included within the text to complement its content. These references have been added to boxes, which list the resource’s URL and a brief description of the resource. All of the source code for the linkedarc.net and related projects has been published to the web as Open-source Software and can be downloaded at http://bitbucket.org/flynam. Finally, the text makes frequent use of technical terms, which are often abbreviated and where this is done, it will be noted in the text. A glossary is included in the appendices, which explains each of these terms.

Conclusions

While this thesis is primarily concerned with the elucidation of a particular technology as it applies to an archaeological problem, I have tried to avoid the trap of techno-fetishism. I see the benefit of using Semantic Web technologies to open up and link archaeological data but at the same time I am not afraid also to criticise these techniques (Huggett 2015). This problem has not emerged out of thin air. It has an historical pretext that manifests itself in the ingrained practices of archaeological institutional bodies and in the minds and methods of the archaeologists that inhabit these social spaces. In that sense, any technological solution that is proposed needs to be acutely aware of these existing needs, assumptions and aspirations. And in a similar vein, these methods will bring with them implications that will be played out in the non-technical realms of the discipline. They will affect the way that archaeological knowledge is produced, disseminated and consumed. They will also alter the discipline’s social reality, in the sense that archaeologists will be required to alter their practice to sit within these new frameworks.

As a final preamble, I want to present one last model for your consideration. This is an important framework, as we will return to it in the concluding chapter when we review the content of the work as a whole and consider what it all ultimately means in the context of our central research question. In an article in which he reflects on the general lack of critical theory in contemporary archaeology towards digital tool use, Huggett suggests that McLuhan’s Laws of Media could be co-opted by the community so as to foster a more explicitly self-reflexive thinking (Huggett 2012; McLuhan 1975, 1977;
McLuhan & McLuhan 1992). McLuhan’s model presents four questions with which a particular medium or tool can be interrogated. These are listed as follows:

1. Amplification: how does this tool enhance or amplify aspects of human function?
2. Obsolescence: what previous tool does its use eclipse?
3. Retrieval: does the use of this tool bring back a previously obsolete activity?
4. Reversal: when used to its fullest potential, does the tool morph into something new?

Huggett illustrates the use of McLuhan’s model with respect to the tool of radio (Huggett 2012 p. 210). Radio enhances access to mass audiences and it renders the print medium obsolete. It also recalls the function of the medieval town crier and when it nears its point of maximum potential, it evolves into the tool of television. Ideally, you will keep these four questions at the back of your mind as you progress through the coming chapters. They should encourage you to question each and every claim and counter-claim made in this work. They should prompt you to ask how each Archaeological Semantic Web technique manifestly affects the practice of real archaeologists. Are these changes brought about for the good of the community or is there some other motivation at play? And ultimately, where might these technological transformations bring archaeology in the next ten or twenty years?


Chapter 2 – Proposing a solution

Introducing the Semantic Web and Linked Open Data as a solution to the problem of heterogeneous archaeological data

‘For a number of years, a small group of archaeologists scattered around the world have been struggling towards what seems an impossible dream: the creation of a world-wide, computer-orientated data bank’ (Chenhall 1971 p. 159)

Introduction

This chapter introduces Linked Open Data and the Semantic Web as solutions to the problem of publishing and accessing heterogeneous archaeological datasets. It begins with a history of the emergence of these two models in the broader socio-technical sphere as responses to a perceived neglect of machine-readable information within the document web paradigm. We then present a technical description of the Linked Open Data model, starting with an overview of the general concepts involved, before moving on to a consideration of their implementation using the Resource Description Framework (RDF) Knowledge Representation (KR) model. Next we look at RDF from a user’s perspective, demonstrating how an RDF triplestore can be queried using the SPARQL query language. This primer on Linked Open Data, RDF and SPARQL prepares the reader for the more advanced technical discussions on ontological modelling found in Chapter 3, and the accounts of the design, employment and consumption of the linkedarc.net and other Archaeological Semantic Web systems included in Chapters 4-6. The chapter concludes with a history of how Linked Open Data and Semantic Web practices have been adopted by the archaeological community.

A history of Linked Open Data and the Semantic Web

The birth of the document web

When two employees of CERN, the world’s largest particle physics laboratory, made their way into a Geneva office on Christmas day 1990 to demonstrate a new information sharing system that they had developed, it is reasonable to assume that they had no idea
as to the chain of events that their presentation would set in train (Connolly 2000).¹⁸ Robert Cailliau and Tim Berners-Lee, who had led the project, wanted to resolve the problem of allowing researchers working at CERN to share information among the community without imposing a unitary formatting standard on that information (Ryan 2010 pp. 106–107). Berners-Lee built a prototype server architecture that would store the user-created content on a central CERN server. He also designed a client that would be able to browse this content. This client software was designed to run on a small number of NeXTStep computer terminals, which he had managed to acquire despite the organisation’s considerable opposition to the project from the outset. These NeXTStep-hosted clients communicated with the server using a new communications protocol called Hypertext Transfer Protocol or HTTP, a fairly basic text-based request-response protocol used to retrieve resources from a waiting server.

If that were the limit of the project’s ambitions, Berners-Lee and Cailliau’s system would now represent only a small mark in the history of the many infrastructures that had been built to share information on in-house institutional networks since the invention of electrical information networks in the post-war years. Berners-Lee, however, decided to appropriate and include in his design the basic logic of a system used to link documents, irrespective of their types, which had first been proposed by the Harvard researcher, Ted Nelson in 1960. He also decided to retain the name that Nelson had originally given the system, ‘hypertext’. He then modified an existing CERN-based document structuring language called SGML to include this hypertext feature and he named this new hybrid, Hypertext Markup Language or HTML (Ryan 2010 p. 107). HTML content operating across a HTTP network came to be known by its two inventors as the World Wide Web, or simply the ‘web’.¹⁹

The web did not, however, have the immediate positive impact that was hoped for by the two men and it languished for a time on the shelves of the CERN offices. Prospects began to improve, however, when an intern hired by Berners-Lee and Cailliau, adapted

¹⁸ This history of the web begins at the point when Robert Cailliau and Tim Berners-Lee enter the stage. However, it is possible to trace the narrative much further back in time, to the code-breaking work of Turing and his colleagues at Bletchley Park in England during the war (Hodges 2014), out of which came the first computer (as we understand that term today), and to the development of the first data networks by the US Department of Defense’s Advanced Research Projects Agency (ARPA) programme (Hunt 2002 p. 2).

¹⁹ The term World Wide Web is credited to Berners-Lee, who according to Cailliau proposed it in the cafeteria of CERN sometime around its 1990 demonstration (Cailliau 1995).
the browser software so as to allow it to be run on computers other than the NeXTStep. Around the same time, a server was set up at the Stanford Linear Accelerator Center in California to house the first collection of hypertext documents outside of CERN. By 1991, with these two additional pieces of the jigsaw in place, the web as the distributed non-hierarchical or decentralised and, most importantly, open information network was born.

This first era of the web, which grew out of Berners-Lee and Cailliau’s work, is now referred to variously as ‘Web 1.0’ or the ‘document web’, and, philosophically, it was inclined towards the promotion of openness and the principle that asymmetric knowledge production and consumption was of value. However, because of the relatively high barriers (finance and knowledge required) to entry, Web 1.0 content tended for the most part to be created by the traditional producers of knowledge. The university sector was one of the first major social institutions to realise the potential of the web for disseminating its content and a large proportion of the first web sites were created and maintained by this academic community.

Another significant milestone in the history of the web occurred when Berners-Lee proposed at a meeting of the ‘www wizards’ the creation of an organisation that would be charged with the technical, strategic and philosophical leadership of the new network. The WWW Consortium, or ‘W3C’ as it came to be known, was established in 1994, and it continues to fulfil this mandate to this day. Many of the technical specifications referred to in this thesis have come out of the efforts of working groups organised by the W3C (W3C 2015).

---

20 It is difficult to isolate exactly when either of these two terms first came into common usage but the suspicion is that they are reasonably recent additions. ‘Web 1.0’ was definitely not used during the time that it refers to. It only makes sense in the context of ‘Web 2.0’, a period of the web and a term that is fairly well defined. The ‘document web’ describes the primary content type that populated this first iteration of the web.

21 An early survey of the experiences of users with browsing content on the web included a list of the top five web sites requested by the sample group (Catledge & Pitkow 1995 p. 4). Ignoring file://localhost requests, of the remaining four sites listed, all were associated with research-based institutions.
Web 2.0

By the final years of the last millennium, changes were beginning to be felt in the top-down information structure of the first web and this evolution was manifested in the introduction of so-called wiki sites (Leuf & Cunningham 2001), which allowed communities of web users to collaborate on the creation of different content types. This gradual change was also seen in the rise of social media sites such as SixDegrees, Friendster, LinkedIn and MySpace (Digital Trends Staff 2014). This movement changed the very nature of the web by allowing any web consumer to become creator and this change was deemed significant enough to warrant the proclamation of a second iteration of the network, known as Web 2.0. Nova Spivack, a man with his finger on the pulse of web trends, defines Web 2.0 as a technology, which focuses ‘on several major themes, including AJAX, social networking, folksonomies, lightweight collaboration, social bookmarking, and media sharing’ (Spivack 2007). While this list encompasses a wide range of activities and processes, some of which are technical and others social in character, Web 2.0’s most radical impact was felt in the types of information that could now be made available. It resulted in an explosion of user-generated content in the form of social media profile pages, images and video.

Although consensus is not a very common phenomenon in discussions of web trajectories, most commentators on the subject would accept that the characteristics of Web 2.0 are still what broadly define the web as it is today. So successful and pervasive has this ‘social web’ model been in fact, that in 2010 Wired magazine was brought to publish its now infamous cover piece, declaring the death of the web (Anderson & Wolff 2010). They reasoned that the web’s users now either spend the majority of their online time frequenting a tiny number of social media sites, most notably Twitter and Facebook, or accessing web content using non-browser software, such as mobile apps.

---

22 A history of such a socially and infrastructurally complex entity as the web is inevitably more involved than this short account might otherwise imply. For instance, it is not entirely true to say that during the Web 1.0 era, all content was being created by large institutions who had the resources to create and curate their own web presences; bulletin boards were allowing these sorts of everyday conversation to happen on the internet since the 1980s (Rheingold 2000). However, these types of social internet service were not strictly speaking part of the history of the web, as they operated using different communications protocols to HTTP such as GOPHER. As such, they can be excluded from our present account.

23 The term ‘Web 2.0’ was first used in a 1999 article written by Darcy DiNucci (1999).

24 Asynchronous JavaScript and XML or AJAX, allows web clients to request content from an online resource without the need to reload the page.
The internet has clearly had an enormous impact on the lives of ordinary people. It is estimated by the Internet Live Stats site that there were in the order of 14 million internet users in 1993 and that most of these would have been using non-www internet systems (internetlivestats.com 2015). This figure rises steadily and dramatically to just under 3 billion users or 40.4% of the world's population by mid-2014 (Figure 1). In reflection of this unprecedented level of growth, a 2011 UN report on the subject of freedom of speech suggested that ‘the internet has become a key means by which individuals can exercise their right to freedom and expression’ (La Rue 2011). A number of media outlets inferred from this that internet access had for all intents and purposes been declared a human right (Kravets 2011).

Open Access, Open Data and the politics and ethics of web access

Berners-Lee and Cailliau’s original concept of the web was of a dynamic and democratic information exchange network. It was to be a place in which all users, irrespective of their financial or social status, would be allowed to operate equally, in the sense that they would all have access to the same content under equal terms and conditions. While the motives behind this idea were highly laudable, realising them was another matter entirely. The first obstacle to achieving these goals was found in the high costs of the technical

---

25 To rate one technology against another is problematic. Is the web the most important technological invention since Gutenberg first started mechanically reproducing text and images in the 13th century, as some would have us believe (Eiteljörg 2011 p. 1)? It is hard to say and, ultimately, we may have to wait for the passing of a number of decades before we can begin to assess the fuller implications (epistemological, ethical, social, et cetera) of the web’s emergence (Olson 2012).
infrastructures required to deliver web access. In the Web 1.0 era, this meant that only certain societal groups (research institutions and state bodies being the most conspicuous) could provide web access to their members. Much has changed in the intervening period, however, and now richer nations have managed to effectively deliver universal web access for all their citizens. The situation is, however, markedly different in the developing world with large swathes of the globe remaining as internet black spots.\footnote{Berners-Lee is now heavily involved with The Alliance for Affordable Internet, which works to address this problem, by promoting cheaper web access in these less developed parts of the world (Gibbs 2013)}

In a related move, web content itself also came under scrutiny. The Net Neutrality movement gives form to the principle that all content on the internet be treated equally, regardless of its source or the way in which the consumer gains access to it. While the term was first used in 2003 by Tim Wu (2003), Net Neutrality really only entered the zeitgeist in May 2014 when FCC chairman, Tom Wheeler set out a plan that would have allowed telecommunications companies to charge content providers for access to faster connections through the internet backbone (Wheeler 2014). Net Neutrality campaigners argued this would have inevitably led to increased costs being passed on to the consumer, thereby leading to the creation of a two-tier web (Goodman 2014).\footnote{It could be argued that internet access has always operated using this model, as end-users pay for the bandwidths between themselves and their ISPs. Charging for the internet backbone lines as well would augment the costing of this existing model but would not constitute a radical restructuring of the charging for access model.} Besides this threat, it is also now commonplace to see the web content’s neutrality compromised by particular political regimes across the world. These governments worry that unbridled access to the web for their people represents a challenge to their political ambitions (Alford 2010). Compounding this is the fact that the web exists within a globalised commercial context and where the possibility of commercial gain exists, there will always be the potential for the original web’s philosophical model to be put under strain (Ryan 2010 pp. 120–136).

The Open Access movement was established to fight against this trend towards a two-tier web in the academic sector. While being compatible with Net Neutrality’s agenda in a broad sense, Open Access is more specific about its objectives. It campaigns against the rise of online academic journal publications being operated for profit by charging their users for content access (Laakso et al. 2011; Lende 2012; Winters 2014). The central
tenets of the Open Access movement state that it is necessary to strive towards a web environment in which all academic content is made freely available online and for this content to be governed by liberal access and re-use licenses (Suber 2012).

Open Data is related to the Open Access model but it is more specific in terms of the content it addresses. It is concerned with the norms that govern the publication of data to the web. While Open Access’s subject is predominantly the journals produced and consumed by the academic community, Open Data’s reach is less restrictive. Its goal is to open up information generated in the civic, political, commercial and research spheres (Kitchin 2014 p. 52). The Open Definition resource provides the following prescription for all forms of open content, including Open Data: ‘Open data and content can be freely used, modified, and shared by anyone for any purpose’ (Open Definition 2015). As such, Open Data emphasises the release of content under liberal access and re-use licenses. As with Open Access, it makes no stipulation as to the form that this data should take.

**On the problem of publishing data to the document web**

Thus far, we have talked almost exclusively about the web as a provider of human-readable content. And we have considered how its enormous potential was realised despite encountering technical, social and political challenges along the way. But this is a thesis concerned with data and with the machine manipulation of this data and, as such, we must now turn our attention to this topic by considering how the seeds of its narrative first took root. It could be argued that the web search engines were the first to use machines to consume the web’s content. They created a vast and lucrative industry out of nothing by using software programs to crawl through the billions of human-readable web pages in order to index them and allow them to be searched by users at extremely high speeds. And while this certainly represents a very large example of the machine consumption of the document web, I would argue that this approach to the document web’s consumption is unnatural, in the sense that these types of web resources were never intended to be interrogated in this way. The ‘bots’ that trawl through the structures that underlie these documents, parsing them line-by-line and ‘scraping’ out their inherent information (Chang et al. 2006; Fernández Villamor et al. 2011) reach their objectives by constructing meaning out of language and the very limited amount of semantic guidance present in their metadata markup.
The problem with this approach is that language can be extremely complicated, particularly when you need to consider multiple languages, as is the case with web content. So, while Natural Language Processing techniques (Jurafsky & Martin 2014; Manning & Schütze 1999) can be applied to overcome some of these difficulties, for the most part, the web search engines are limited in terms of the meaning that they can recover from human-readable resources. It was never intended that HTML be consumed in this way. Its raison d’être is to communicate a document’s structure to a client, normally a web browser, which then uses this information in order to display the content that it encompasses. While HTML does contain a degree of semantic information about its subject (for example, the inclusion of document elements such as headers, footers, paragraphs and headings), for the most part, it has very little to say about the deeper semantics of the document’s content.

There is also the matter that as users of non-web based computer technology, we use file formats that are not part of the traditional document web canon. For this project, we are interested in the types of ‘offline’ documents that manage information. We have already introduced these file types in Chapter 1 as Archaeological Information Systems. Can these documents also be published to the document web and, if they are, how can they be found and accessed by other users? For example, how does one go about relating the data contained within one online database to another? Can they, for instance, also include hyperlinks as in the hypertext model? As archaeologists, we encounter FileMaker databases regularly. What does it mean to publish data in this format to the web? Does it make its content indexable by the search engines? Is the context of each of its units of information visible to the web consumer?

Information Systems come in a variety of forms but the majority are built upon the relational model and are known as Relational Database Management Systems or RDBMSs. An RDBMS is an information resource, which is separated into different categories of information, which are known as tables. These tables are two-dimensional information stores, in which each row or record contains further sub-units of information, known as fields. The RDBMS is relational because each record can be and usually is conceptually related to records within other tables of the same RDBMS. The
RDBMS model has proved enormously successful and popular as a structure for information storage in the digital age, particularly in its SQL manifestation.

The RDBMS is useful not only because it allows information to be rationally organised but also, and perhaps principally because, it allows that information to be retrieved systematically. This model has worked well for RDBMSs when the objective is to query a single database unit. For example, it is a trivial matter to find a piece of information contained within a single FileMaker database, an Excel spreadsheet, an Access database or a SQL database. Problems begin to emerge, however, once you think of the RDBMS within a distributed environment. How is the information contained within one RDBMS to be linked with other RDBMS information? How is all this RDBMS data to be accessed and interrogated by a user? For the standard documents associated with Web 1.0 and Web 2.0 publications, there is no need to use any software other than a web browser to view or read their contents. For example, an HTML page, an image and a video can all be viewed using a web browser. Files that contain structured data are a different matter, however, and most browsers will not allow for this content to be viewed from within their native user interface. In these cases, the particular resource, whether it is an Excel spreadsheet file, FileMaker database or some other non-standard data resource, will need to be downloaded first and then opened using software that can read the file format. Often, these formats are proprietary and the user needs to acquire specialised software to access the contents.

Indexing of online RDBMS resources also presents a problem. While Google, the most omnipresent of the search engines, states that it does index certain spreadsheet documents (Google 2015a), a search query that targets data inside an Excel document for instance, will not give the user any further information about the nature or context of that match. And it is this context that is valuable to the consumer. The simple assertion that a textual match has been made is of limited use. Another issue that comes up with the use of RDBMSs in a web context is that the RDBMS is designed to be self-contained and this means that they are entirely ignorant of the web resources that surround them. This picture of isolation has led online RDBMSs to attract the pejorative moniker of ‘data silo’ or even ‘data tomb’ (Gattiglia 2015 p. 1). Online data silos fail to profit from

---

28 As an historical aside, the first evidence of the use of a two-dimensional table to represent the relationships between information entities is seen in an analysis of differential soil temperatures published by Johann Heinrich Lambert in 1779 (Friendly & Denis 2001 p. 13).
the fact that they exist within a networked environment, which could potentially offer resources that would be of use to the data silo’s contents.

Latterly, there has been a push away from the RDBMS model in favour of a less structured model called a ‘document database’ or (and in reference to its hegemony in the sector) ‘NoSQL’. These systems organise their data in a different way (key-pair, document and graph are the dominant approaches) to the tabular structure of the RDBMS. This more flexible model allows for the storage of a more diverse range of information types and for greater flexibility in terms of scale. It remains to be seen if these potentially more transparent data frameworks will realise this potentiality in terms of their visibility to crawlers of web content and their reliance on other web data resources.²⁹

**Introducing Linked Data and the Semantic Web**

linkedarc.net resource

An interactive timeline of the key publications in the history of the Semantic Web

http://linkedarc.net/datamining/lodlam-zotero

In 2001, Tim Berners-Lee, James Hendler and Ora Lassila (2001) outlined a grand vision for the resolution of these web data problems, which they called the Semantic Web.³⁰ In many respects, their paper was written from a position of power. The document web had proved to be an enormous success for the digital publishing of human-readable content and now it was time to do the same for data. In hindsight, it is easy to spot a suggestion of hubris in the approaches and outcomes that the paper describes and predicts. The paper opens with a fictional account of a use of the Semantic Web. This scenario sees Pete and Lucy work to resolve a medical dilemma involving their mother. In fact (and this is the point of the introduction), Lucy and Pete do very little themselves to find a solution to their problem. Instead, they merely input a set of parameters into their semi-intelligent web-connected devices or ‘agents’ and the AI agents figure out the solution. In a matter of minutes, the agents have suggested a plan of action for the scheduling of physical therapy sessions for Lucy and Pete’s mother. They accomplish this by

---

²⁹ We discuss the pros and cons of using FileMaker, SQL and NoSQL to manage archaeological digital data in Chapter 7.

³⁰ The Semantic Web is also often referred to as the Web of Data.
constructing queries based on Lucy and Pete’s preferences and using these queries to interrogate the sum total of the web’s knowledge, which, it is inferred by the paper, is great indeed.

The objective of this fictional preface is to paint an idealised view of what the web could become, were it to be improved upon. This ‘Semantic Web’ would allow machines to extract the semantic content out of web pages. As was explained in the previous section, the information needed to answer the problem of scheduling Lucy and Pete’s mother’s physical therapy session is already available on the web but its current representation makes it impossible, or at least very difficult, for it to be retrieved without the involvement of a human agent. Berners-Lee et al emphasise that there would be no need to create an entirely new stack of technologies to realise this Semantic Web, as it would operate as an extension or an ‘evolution’ of the existing web. As such, it would sit on top of the same core infrastructural components (HTTP communication, URI addressing, hyperlinks, HTML structure). This backward compatibility aspect of the new web was to become one of its key selling points in the years to come.

Despite the exuberance of its tone and the fact that the 2001 paper essentially outlined all of the components needed to achieve the Semantic Web, the years following its publication did not see any major Semantic Web developments at a grass roots level. This led Berners-Lee to publish a follow-on paper, *Linked Data – Design Issues*, in 2006 (Berners-Lee 2006). Isaksen (2011 p. 28) speculates that Berners-Lee’s decision to publish the second major paper at this point was due to his exasperation at the lack of progress in the form of real world implementations of Semantic Web principles. Whatever the reason, this second milestone paper can be read with a more practical agenda in mind. While the 2001 *Semantic Web* is written in an aspirational style, *Linked Data – Design Issues* is more rooted in the here and the now of the matter and with the prosaic objective of encouraging more people to start producing data compatible with the Semantic Web model.

Careful reading of the two papers does, however, show signs of a certain movement in Berners-Lee’s approach to the subject. The 2001 document focuses primarily on the need to populate this new web with semantically rich content. It speaks of the need to extensively employ the as yet untested promise of Knowledge Representation (KR). KR
Proposing a solution

40

is a sub-field of Artificial Intelligence, which is concerned with the serialising of digital content in such a way as to allow its comprehension by machines. These representations need to capture formal reasoning rules, which can then be employed by the machine to assess the semantic value of the data that it encounters. Systems, known as ontologies, are to group together the implementations of these rules so that information from one source can be rendered comparable to information from another.31

The objective, therefore, of the 2001 article is to provide for an environment in which AI agents can carry out specific tasks and that this will ultimately serve to ‘assist the evolution of human knowledge as a whole’ (Berners-Lee et al. 2001). This ambition is offset against the pragmatism of the 2006 article, which emphasises more explicitly the need to link information together in the same way that HyperText allows documents to be linked. Linking data, the paper argues, will bring about more data discoverability. This will make the Semantic Web more valuable and as a result, it should attract a greater uptake. The abstract talk of the first paper is discarded in favour of the use of examples. Ontologies are referenced (in the form of FOAF and RDFS32) but only as a means to a practical end and less in the epistemological sense that they are referred to in the first paper. The mechanics of a Semantic Web conversation are also presented, again in a practical sense, with Berners-Lee outlining how a HTTP server should respond to various forms of URI requests.

In 2010, Berners-Lee made an amendment to the 2006 paper by posing the question: Is your Linked Open Data 5 star? (Berners-Lee 2010) In this addition, Berners-Lee draws a distinction between Linked Data on the one hand and Linked Open Data on the other. In the intervening four years Berners-Lee obviously felt that stressing the ‘openness’ of the dataset was important in the context of its fitting into the Semantic Web model and that the Semantic Web as an Open Data system needed to be made explicit. As such, the Linked Open Data annex talks about the importance of using open standards for data representation (RDF) and querying (SPARQL). It also reinforces the general tenet that Semantic Web content should be structured, albeit remaining flexible within this structure. With the 2010 change, the paper essentially becomes a checklist by which any

31 Chapter 3 includes a detailed explanation of the ontology’s role in the Semantic Web model.
32 Friend of a Friend (FOAF) is an ontology used to model people, organisations and their relationships. It is discussed further in Appendix A. Resource Description Framework Schema (RDFS) is an ontology used to structure other ontologies. It is discussed in Chapter 3.
dataset can be rated for its adherence to Linked Open Data practice and perhaps because of this simplicity of purpose it has become a fundamental go-to source for the subject.

Figure 2: the Linked Open Data cloud as of 30 August 2014

The years following the publication of *Linked Data – Design Issues* have seen further policies and technical approaches being proposed but, for the most part, the parameters of the field have remained largely as set out in this pair of Berners-Lee articles. The obvious question that presents itself in the light of these two documents is, did they have an impact at grassroots level? Did they serve to entice more data providers into producing Linked Open Data? To attempt to answer this question, one could do worse than referring to the Linked Open Data Project, which was founded in 2007 to ‘bootstrap the Web of Data by identifying existing data sets that are available under open licenses, converting these to RDF according to the Linked Data principles, and publishing them on the Web’ (Bizer et al. 2009). The project periodically updates a Linked Open Data ‘cloud graphic’, which shows the inter-connections of all registered Linked Open Data datasets. The latest version of this is reproduced in Figure 2.

The project has also conducted two studies into the state of the sector (published in 2011 and 2014) and the latter of these (Schmachtenberg et al. 2014) suggests the emergence of some interesting trends in the last number of years. In overall terms, the amount of the
Linked Open Data datasets has grown by 271% in the three years that separate the two studies, reaching a figure of 1,091 in 2014. The Linked Open Data Project report does not give a number for the amount of information that this equates to but in 2011 the total size for the 294 datasets is listed as being over 31 billion RDF triples.\(^{33}\) The LODStats project (Ermilov et al. 2013) does, however, provide a more recent figure\(^{34}\) and while its list of source datasets is different from that of the Linked Open Data Project reports, its figure of nearly 90 billion RDF triples does, nonetheless, largely tally with what would be expected for the Linked Open Data Project second report. Given that there were just 12 compliant datasets listed when the project first started collecting this information in 2007, this is a sign of the sector’s progress at least in terms of raw numbers.

The Linked Open Data Project 2014 report also looked at levels of adherence to Linked Open Data best practice and these trends are more mixed. The percentage of datasets that link to external datasets is set at 56%, a figure which is down from about 70% in the 2011 report.\(^{35}\) There is also a fall in the rates at which datasets are making their data available using more than one interface (in 2014 only 9.96% advertise SPARQL endpoints and 8.19% provide their data as a raw dump, down from 68.14% and 39.66% respectively in 2011). And there are also mixed results for the publication of basic dataset metadata – in 2014 39.69% of datasets provided provenance information and 9.96% licensing details, which compares against 20.41% for provenance and 17.84% for licensing details in 2011.

The types of dataset being published as Linked Open Data have changed in this period, as well. Social media datasets have gone up with most of these being represented by small FOAF-based publications, often associated with outputs produced by plugins for platforms such as WordPress (Bizer 2009 p. 90). Datasets associated with government, publications, life sciences, user-generated and cross-domain subjects have all also grown with just geographic and media datasets falling slightly.

\(^{33}\) An RDF triple is the basic unit of information within the Linked Open Data ecosystem. A full description of the RDF triple is provided later in this chapter.

\(^{34}\) Its website states that it was last updated less than 2 months ago, as of 15 July 2015 (AKSW 2015).

\(^{35}\) Having said that, it needs to be borne in mind that the two reports are not exactly comparing like with like, because the 2011 study seems to have imposed stricter compliance criteria on its participants than the 2014 study, and this discrepancy needs to be taken into account for all of the comparisons included here (Schmachtenberg et al. 2014 p. 14).
Before 2007 there was no DBpedia (Lehmann et al. 2014), Wikidata (Vrandecic 2013) and YAGO (Hoffart et al. 2013; Mahdisoltani et al. 2014), but all of these now dominate or are set to dominate the field in the years to come in terms of the amount of data that they hold and their role as central link hubs within the ecosystem. The Google-managed Freebase, which was founded in March 2007 and is another of these hub-like cross-domain data providers, is set to close in the coming months,\(^{36}\) and its data to merge with that of Wikidata. This would allow Wikidata to perhaps challenge DBpedia’s dominance in the future. While it looked like the 2011 Linked Open Data cloud was becoming somewhat centralised with these big players acting as magnets for smaller datasets, the Linked Open Data Project 2014 report concludes that this trend is now correcting itself and that the current Linked Open Data cloud is more evenly distributed between small and large data projects. So, in that sense, the idea of the Semantic Web as a decentralised knowledge domain seems to be gradually coming to pass.

**Is Semantic Web data Big Data?**

Before we move on, it is worth considering briefly the subject of Big Data, not because it is the focus of this work but, because it is currently the topic du jour and its relationship, if any, to the Semantic Web needs to be clarified. Big Data has attracted various definitions, which have been summarised by Diebold (2012) and Ward and Barker (2013). The term itself was coined by John Massey, a Silicon Valley executive in the mid-1990s. Most definitions focus on the variable of quantity. A 2001 Meta report (Laney 2001), despite not mentioning the term Big Data itself, cites the importance of the three Vs – Volume, Velocity and Variety – in identifying such data. Intel goes along with this scale principle, stating that Big Data organisations need to be producing ‘a median of 300 terabytes of data weekly’, and that this data can be unstructured and is intended to be subjected to analytical processing (Intel IT Center 2012). Oracle follows this latter perspective, eschewing the focus on data quantity in favour of a definition that is based on the types of data involved, with Big Data including more unstructured information coming from nontraditional sources, such as blogs, and other social media, and the storage infrastructures and methodologies employed to manage and to gain value from that data (Dijcks 2012). Oracle also highlights how in their understanding Big Data datasets tend to be associated with the latest generation of RDBMS systems, the so-called NoSQL solutions, introduced earlier in this chapter. Microsoft also considers Big

---

\(^{36}\) As of July 2015.
Data from this methodological viewpoint, foregrounding the computer’s role (including Artificial Intelligence techniques) in Big Data processing (Gattiglia 2015 p. 2; Microsoft News Center 2013).

In terms of technological platforms, currently the most ubiquitous Big Data application framework is the Apache Hadoop platform (Tibco 2013), which was designed to deliver ‘reliable, scalable distributed computing’ (The Apache Software Foundation 2014a) that can be used to host extremely large datasets. Hadoop has been criticised for not being designed with the average user in mind, that it is far too complex and abstruse for all but the most committed and technically savvy user groups to master. On the other hand, its abilities to analyse and visualise enormous datasets are heralded as being second to none.

My own initial view of Big Data was that it defined datasets, which are complex and which demand the use of specific computational methodologies if meaning is to be successfully extracted from them. Or, as Robert Hillard of The Method for an Integrated Knowledge Environment (MIKE2.0) project puts it, ‘Big Data can be very small and not all large datasets are big’ (Hillard 2012). In this understanding, Semantic Web data would qualify as Big Data. However, with further consideration, I have changed this view slightly. I now see the two primary aspects that set a Big Data dataset apart from other datasets are volume and variety. On the first point, Semantic Web data might in certain circumstances qualify – DBpedia’s datastore, for instance currently hold over 1 billion RDF triples. However, for the most part, the sizes of dataset that are encountered would not justify their labelling as Big Data.37 On the variable of structure, RDF, as we will see in the next section of this chapter, is highly structured and I believe that this sets it apart from the inherent lack of structure that typifies the Big Data dataset. For both of these reasons, Semantic Web content, particularly that which can be termed archaeological in nature, should not be considered as Big Data.

---

37 The Priniatikos Pyrgos RDF dataset, which was created as part of this research project and is described in Chapters 4-6 of this thesis, as of 15 July 2015, holds just over 4 million RDF triples. This could not be viewed as Big Data in terms of volume.
A note on terminology and trends

Figure 3: the frequency of Google searches related to Linked Open Data and the Semantic Web since 2004 (Google 2015b)

As with most fields, the use of terminology surrounding the Semantic Web has evolved over the years. Figure 3 shows a graph created using Google Trends data, which charts the relative frequencies of the use of a number of terms mentioned in this chapter as Google search queries against time from 2004 to the present day.38 A clear pattern can be identified in the decline of the use of the term Semantic Web from its height in 2004 to the present day. Open Data, on the other hand, goes in the other direction, growing in popularity as a search term over the same period, while Open Access falls from a position of dominance during the 2004-2007 period; before steadying out to remain largely unchanged from that point on. Linked Data and the related Linked Open Data on the other hand are much less significant when compared against these other terms. It is also informative to note that Linked Open Data first appears on Google’s radar in late 2010, a date which is significantly later than the other terms and which is in keeping with its first use by Berners-Lee (Berners-Lee 2010).

---

38 Note that the use of some of these terms by users of Google’s search engine predates 2004. However, the Google Trends service only provides data from this point in time on.
Given that we have talked at length in this chapter about the web in terms of its evolution as first Web 1.0 and then Web 2.0, it is perhaps unsurprising that a number of commentators (Hendler 2009; Lassila & Hendler 2007; Markoff 2006; Spivack 2007), have called for the current and future direction of the web, including its semantic extension, to be labelled web 3.0. Spivack’s definition of the key aspects of each phase of the web is summarised in graphical form in Figure 4, while Figure 5 compares the Google search frequencies for both Web 2.0 and web 3.0. Interestingly, the use of both

---

39 See Barassi and Treré (2012) for an argument against this practice of dividing up the development of the web into separate technological and social phases.
terms seems to be falling, with there being a marked decline for Web 2.0, which is coming from a high in 2007. The decline is more gradual for web 3.0, which has yet to really take off in the public consciousness.

As Willard McCarty remarked when he set out the terms for his investigation into the field of Humanities Computing (McCarty 2005 p. 2), the point of terminological explicitness is not to demarcate or limit the extent of the coming discussion, but merely to remind the reader of the historical nature of the particular discourse. For the purposes of this thesis, from this point on, the term Linked Open Data will be used to refer to a Knowledge Representation, which is published to a digital network, which need not be a part of the web, under an open license and which is composed of individual data nodes. These nodes can be linked to form a graph of networked information. The Semantic Web on the other hand refers to the collection of Linked Open Data resources, which is published to the web and whose structure is made explicit with the use of published ontologies.

**A technical primer on Linked Open Data and SPARQL**

-linkedarc.net resource

A series of introductory videos to topics such as Linked Open Data and SPARQL

http://linkedarc.net/about/movies

The preceding section outlined the historical events that saw the birth of the Semantic Web, first as an aspiration in the late 1990s and then as a call to grassroots action in the 2000s. While the key concepts of the paradigm have been introduced, we have yet to talk about it in explicitly technical terms, which is a key requirement if you are to be able to understand the contents of the following chapters. As such, the purpose of this section is to first outline the principles of the Linked Open Data KR philosophy and its implementation, Resource Description Framework (RDF). We will then turn our attention to Simple Protocol and RDF Query Language (SPARQL) and devote some time to explaining how SPARQL can be used as an effective tool with which to query Semantic Web content.
Linked Open Data and RDF broken down

There are a number of sources that give technical overviews of Linked Open Data but one could do far worse than returning to Berners-Lee’s 2010 updated *Linked Data – Design Issues* article for initial guidance (Berners-Lee 2010). In *Linked Data – Design Issues*, Berners-Lee sets out four rules that a dataset needs to comply with, if it is to be considered Linked Data.

The first requirement is that every thing or every entity that is manifest in the dataset needs to be identified using a Uniform Resource Identifier or URI.40 A URI is simply a unique identifier associated with a resource. For example, if your dataset contains a digital record that represents a dagger artefact, which was found at a site in Ireland, then the digital record needs to be given an address along the lines of ‘datasetflynam-artefact-dagger1’ and the site might be given an address of ‘datasetflynam-ireland-site1’. The important thing to remember is that these identifiers should be unique not only to your dataset but also to all datasets. This is a simple idea but it has very powerful implications in the context of data sharing, as we will see.

Second, these URIs should conform to the HTTP format for URIs. This format is the same as the Uniform Resource Locator (URL) format. URLs are the addresses used to identify web resources (Segaran et al. 2009 pp. 64–66). The URIs that we suggested in the example described above could be transformed into `<http://example.org/dagger1>` and `<http://example.org/site1>`.

Using HTTP addresses as URIs for data also helps resolve the problem of ensuring that the label used is globally unique, as domain names once purchased cannot be used by any other data provider.

The third Linked Data rule states that these URIs should be dereferenceable, which means they must be associated with a tangible web resource. The simplest example of this principle in action is the retrieval of a web page using a browser. The user enters the URL of the resource into their browser’s navigation field and the browser retrieves and

---

40 The Digital Object Identifier (DOI) is sometimes used in the place of the URI, and it is referenced on occasion by this thesis. While the DOI is similar to the URI, it is not exactly the same. While anyone can create a URI, as long as they own the base domain of the URI, DOIs must be allocated by certified DOI Registration Agencies. While not required, modern DOIs often follow the HTTP addressing scheme used by URIs.

41 The example.org and example.com domains are special domains that are intended to be used in examples such as this. They are reserved domains, which cannot be purchased.
displays the web page and any other resources that are associated with the URL. As such, URLs are a sub-class of URIs. Another aspect of the third rule is that the web engine, which serves these resources for the URI in question should provide ‘useful information’ and that this information should be formed ‘using the standards’. The standards that Berners-Lee explicitly mentions are RDF and SPARQL. Finally, the fourth rule of Linked Data specifies that this RDF data should link to the URIs of other RDF data existing on the Semantic Web.

The 2010 version of the *Linked Data – Design Issues* paper expands the Linked Data model to include Open Data, thereby creating Linked Open Data. The difference between Linked Data and Linked Open Data is the fact that Linked Open Data must also use open standards (RDF and SPARQL already qualify as such) and employ liberal access and re-use licenses to govern the data that they hold. In that sense, Linked Open Data publications also need to explicitly state the licenses, which govern the access and re-use of their data. Over the years, a number of groups such as Open Data Commons (2015) and Creative Commons (2015) have made available a range of license types to meet the various licensing needs of data providers (*ARIADNE 2014a* pp. 11–13; Zuiderwijk & Janssen 2014).

Something that needs to be understood about Linked Open Data is that in itself it is merely a KR philosophy or a set of guidelines for best practice. On its own, it is not of much practical use. It needs an implementation to be realised and by far the most common approach to applying the Linked Open Data rule-set is Resource Description Framework or RDF. RDF first appeared as a W3C specification recommendation in 1999 (Lassila & Swick 1999). This was followed with an update in 2014 (Cyganiak et al. 2014) and, as we have seen, it has enjoyed a close working relationship with Semantic Web applications ever since. RDF is an extremely straightforward model, which allows a computer to represent a relationship between two entities (Antoniou & Van Harmelen 2004 p. 62). This representation, known as an RDF triple, is made up of three elements, a subject, an object and a predicate, which links the subject to the object as in Figure 6.
Let us consider a simple example, which encodes the meaning represented in the following sentence.

‘Context 101 is a cut’

This can be represented schematically as an RDF triple as is shown in Figure 7.

While not a requirement of RDF, when used in a Semantic Web context, it is common practice for each of the triple elements to be represented as a URI. The URIs chosen will depend on the data system being developed. Usually, a data provider will reserve the use of a HTTP URI domain by purchasing it, as this allows the base domain to be used to construct each of the URIs needed to satisfy the identification requirements of the dataset, as is shown in Figure 8.

In the **RDF 1.1 Concepts and Abstract Syntax** (2014 p. 1), RDF now allows for the use of Internationalized Resource Identifiers (IRIs), which are a generalised form of URIs, in that they contain all of the features of the URI, as well as an ability to be expressed using Unicode characters.
In and of itself, a triple is limited in terms of the level of semantic complexity that it can model. However, the ability to chain RDF triples together by using one triple’s object as the subject of a second triple creates what is known as a RDF graph. And this can theoretically encode very complex and substantial meanings (Figure 9).

Any of the three elements that make up an RDF triple can also be represented in the form of what are known as literals, but in practice literals are mostly reserved for use as objects. A literal is a computing term used to define a fixed value and in the case of RDF this is the set of all fixed values excluding URIs. Literals come in different forms depending on the type of data being represented. The most common type used in RDF is undoubtedly the text value (known as a string in computing) but literals can also represent integer or floating-point numbers and dates (Peterson et al. 2012). The use of a literal as an object value constitutes the end of the line for the data graph in the sense that a literal cannot link to another URI via a predicate because a literal is not a URI.
However, literals are extensively used within RDF graph data as they allow the data producer to populate the dataset with facts. An illustration of this, which extends the Context 101 example cited above, is represented in Figure 10.

![Figure 10: the Context 101 RDF triple example including a literal](image)

RDF is a conceptual framework. In order to store RDF on a computer, it needs to be committed to disk using a particular format and we call this its serialisation. Serialisation is analogous to the act of writing and in the same way that the same language meaning can be written using different scripts, it is also possible to serialise the one RDF meaning using different serialisation formats. The simplest form of RDF serialisation is known as N-Triples (Berners-Lee 2005) and an example of this is shown in Table 1. This serialisation encodes the meaning graphically illustrated in Figure 10.

<table>
<thead>
<tr>
<th>Table 1: an N-Triple example</th>
</tr>
</thead>
</table>

The format of N-Triples is extremely simple. Each triple is entered as a subject, predicate and an object, with each element being separated by a space. The triple is terminated by a period. URIs are enclosed in angled brackets and literals in double-quotation marks. It is customary, although not a rule, to enter a triple on its own line. One of the best features
of N-Triples and in fact of all RDF serialisations is that they can be created using any basic text editor software.

You may have noticed that of the two predicates used in the example, neither address is derived from the example.org domain. These predicates, `<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>` and `<http://www.w3.org/2000/01/rdf-schema#label>` have in fact been created by someone else and are published to the web as public ontologies. This renders them available for reuse by other RDF datasets. We will cover the subject of ontologies at some length in the next chapter but for now it is sufficient to be aware that we are making use of an external resource that gives the affected RDF triples a particular meaning.

**RDF/XML**
```xml
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF
  xmlns:rdfs="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:rdf="http://www.w3.org/2000/01/rdf-schema#"
>
  <rdfs:Class rdf:about="http://example.org/context101">
    <rdfs:label>Context 101</rdfs:label>
  </rdfs:Class>
</rdf:RDF>
```

**Notation3**
```xml
@prefix dc: <http://purl.org/dc/elements/1.1/>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix xsd: <http://www.w3.org/2000/01/XMLSchema#>.

<http://example.org/context101> a <http://example.org/ContextTypeCut> ;
  rdfs:label "Context 101".
<http://example.org/ContextTypeCut> dc:creator <http://example.org/Franklynam>.
```

**JSON-LD**
```
{
  "@context": {
    "dc": "http://purl.org/dc/elements/1.1/",
    "rdf": "http://www.w3.org/1999/02/22-rdf-syntax-ns#",
    "rdfs": "http://www.w3.org/2000/01/rdf-schema#",
    "xsd": "http://www.w3.org/2001/XMLSchema#"
  },
  "@graph": {
    "@id": "http://example.org/context101",
    "@type": "http://example.org/ContextTypeCut",
    "rdfs:label": "Context 101",
    "dc:creator": {
      "@id": "http://example.org/Franklynam"
    }
  }
}
```

Figure 11: different RDF serialisations all encoding the same meaning

Besides N-Triples, RDF can be serialised in a number of different ways. RDF/XML was the first RDF serialisation and remains one of the most common forms encountered (Gandon & Schreiber 2014). There is also the option to use N-Quads (Carothers 2014), which extend the N-Triples functionality, Turtle (Beckett et al. 2014) and Notation3 (Berners-Lee et al. 2008). Recently, the popular web serialisation format, JSON has been extended to support RDF data (Sporny et al. 2014). RDF data can even be embedded...
into the HTML of webpages using RDFa (Adida et al. 2015). The point to remember is that choosing to use one serialisation over another does not change the meaning that the serialisation encodes. All of the serialisations shown in Figure 11 have identical semantics.

For the purposes of this thesis, almost all RDF examples will be presented using the Turtle serialisation. The reasons for choosing Turtle are because it is very easy to read and it also includes certain functionality (prefixes and triple clustering) that renders it more concise than many of the other serialisations. In practice, most RDF datastores (known as triplestores) allow for users to request information in most if not all of the available RDF serialisation formats. This means that, if you choose to devote your efforts to becoming proficient in the writing of one format and not another, this will not limit the potential future uses that the RDF data can be put to in any substantive way.43

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>@prefix dct: <a href="http://purl.org/dc/terms/">http://purl.org/dc/terms/</a> .</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>@prefix rdf: <a href="http://www.w3.org/1999/02/22-rdf-syntax-ns#">http://www.w3.org/1999/02/22-rdf-syntax-ns#</a> .</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>@prefix rdfs: <a href="http://www.w3.org/2000/01/rdf-schema#">http://www.w3.org/2000/01/rdf-schema#</a> .</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>@prefix example: <a href="http://example.org/">http://example.org/</a> .</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>@prefix foaf: <a href="http://xmlns.com/foaf/0.1/">http://xmlns.com/foaf/0.1/</a> .</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>@prefix xsd: <a href="http://www.w3.org/2001/XMLSchema#">http://www.w3.org/2001/XMLSchema#</a> .</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>example:context101 a example:ContextTypeCut ;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>example:ContextTypeCut dct:creator example:FrankLynam .</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>example:FrankLynam a foaf:Person ;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>foaf:name &quot;Frank Lynam&quot; .</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: an example RDF Turtle serialisation

An example RDF Turtle serialisation is shown in Table 2.44 It extends the Context 101 example by encoding a more extensive set of meanings associated with the context. Turtle allows you to shorten the URIs used throughout a document by making use of its prefix system. In this example there are six prefixes declared on lines 1-6. By declaring,

43 In any case, it is a relatively trivial matter to convert from one RDF serialisation to another and there are a number of tools that can help you to do this (Humfrey 2015; Stolz et al. 2013).

44 Note that the use of tabs and spacing in this example is purely to make the RDF appear more readable. They do not change the semantics of the RDF content.
for instance, a prefix ‘dct’ for the <http://purl.org/dc/terms/> URI base, it is then possible to enter dct:created within the triple serialisations in place of the more verbose <http://purl.org/dc/terms/created>.

RDF Turtle’s syntax allows for the clustering of triples that share the same subject. This is exemplified on lines 8-12, which serialise a set of triples that all have example:context101 as their subject. Instead of re-entering example:context101 for each of these triple definitions, it is possible to shorten the syntax using the semicolon special character. Therefore, line 9 serialises the triple example:context101--rdfs:label--"Context 101" and so on. Another special piece of syntax contained in these triples is the ‘a’ keyword. This has the same meaning as entering rdf:type. rdf:type is one of the most commonly used predicates in Linked Open Data datasets, which should not come as a surprise given that it is the accepted way of assigning a type to an object. Assigning types is a fundamental step in making a dataset conform to an ontology and we will learn more about this topic in Chapter 3.

Lines 9, 10 and 11 all serve to associate the subject example:context101 with a series of object literal values. The first two of these literals are in the form of strings, while the final literal is a date value. Dates and other non-string literals are declared as values contained within double quotation marks with the addition of a postfix specifying the value’s type. These various types are declared within another public ontology, xsd.

Finally, the Turtle file declares in lines 14-17 two more groups of triples, the first of which serialises the meaning for the example:ContextTypeCut resource and the second for the example:FrankLynam resource. Note that both of these have been linked to by triples contained within the example:context101 triple grouping. If we wanted to publish this data, we would merely need to save it to a file and then upload that file to the web server paired with the resources’ domain, http://example.org. We would finally publish an additional metadata file outlining the data’s licensing conditions and provenance, and at that point the dataset would have satisfied all of its commitments, thereby making it a legitimate part of the Semantic Web.

45 A recent study by Hees (2015), which analysed a number of different aspects of DBpedia, showed that rdf:type, at 86,391,520 instances, was the second most frequently used predicate after dbpedia-owl:wikiPageWikiLink, which came in at 149,707,899 uses.

46 See the comment above about the inability to use example.org in real world data scenarios.
In this admittedly condensed introduction to RDF Turtle, you will have learnt enough about RDF serialisations to allow you to read the RDF examples contained within the rest of this work. Before continuing, it is worth reviewing the content that has been covered by subjecting our Context 101 example to the Berners-Lee 5 Stars of Linked Open Data rating checklist. We can award our dataset one star because we have published it online under an open license. As it is encoded as RDF, we can also say that it is machine-readable, which gives it two stars. RDF is also a non-proprietary (three stars) and an open-standard format (four stars). Finally, deciding whether or not to award the last star to our dataset is slightly less clear. Our dataset does link to external datasets because it makes use of the public DC Terms, FOAF, RDF and RDFS ontologies but in my opinion this is not exactly what Berners-Lee had in mind when he set the criteria for this last star. It is far more likely that his prototypical 5-star Linked Open Data dataset is a network of data, which is as much embedded within the fabric of the greater Semantic Web, as an HTML document on a website is embedded in the document web. So, in order to achieve this final star, we really need to include external links as object values in our graph data. In practice, this is where most aspiring Linked Open Data datasets fall down because they simply do not trust that these external resources will be available when required (Buil-Aranda, Hogan, et al. 2013) and even if they can overcome this issue of trust, it is unclear how these external resources might be explored using the current RDF technological stack.47

**RDF triplestores and the SPARQL way**

linkedarc.net resource

A practical introduction to writing SPARQL queries addressed to linkedarc.net

http://linkedarc.net/about/sparql

The previous section described the structure of RDF and how RDF can be created. Now it is time to consider the RDF system from the perspective of the consumer. In this section we will introduce SPARQL as, I would argue, the most flexible and powerful approach to extracting specific meaning from within an RDF triplestore. SPARQL was

---

47 This matter of trust and the lack of a standard technical infrastructure to allow for querying across more than one RDF dataset are major obstacles to the growth of the Semantic Web and we return to both in Chapters 7 and 8.
first proposed by the W3C in 2008 (Prud’hommeaux & Seaborne 2008) and an updated specification, SPARQL 1.1, was released in 2013 (Buil-Aranda, Corby, et al. 2013). To date, the amount of active and maintained SPARQL endpoints on the Semantic Web is very low – on 2014 estimates, 9.96% of RDF datasets provide access to their data via a SPARQL endpoint (Schmachtenberg et al. 2014 p. 14). The level of support by RDF datasets dealing specifically with cultural heritage and archaeological content might be slightly higher than this figure, with Isaksen (2011 p. 59) noting that 24 of 70 projects (34%) surveyed in his 2011 survey of Linked Open Data practice within the archaeology and cultural heritage fields stating that they provided support for SPARQL querying.

Table 3: an example RDF Turtle serialisation

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>rdfs:</td>
<td><a href="http://www.w3.org/1999/02/22-rdf-syntax-ns#">http://www.w3.org/1999/02/22-rdf-syntax-ns#</a> .</td>
</tr>
<tr>
<td>rdfs:</td>
<td><a href="http://www.w3.org/2000/01/rdf-schema#">http://www.w3.org/2000/01/rdf-schema#</a> .</td>
</tr>
<tr>
<td>example:</td>
<td><a href="http://example.org/">http://example.org/</a> .</td>
</tr>
<tr>
<td>foaf:</td>
<td><a href="http://xmlns.com/foaf/0.1/">http://xmlns.com/foaf/0.1/</a> .</td>
</tr>
<tr>
<td>xsd:</td>
<td><a href="http://www.w3.org/2001/XMLSchema#">http://www.w3.org/2001/XMLSchema#</a> .</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RDF Turtle Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>example:context101 a example:ContextTypeCut ;</td>
</tr>
<tr>
<td>rdfs:label &quot;Context 101&quot; ;</td>
</tr>
<tr>
<td>rdfs:comment &quot;This is the cut associated with fill Context 100&quot; ;</td>
</tr>
<tr>
<td>dct:created &quot;2008-07-24&quot;^^xsd:date ;</td>
</tr>
<tr>
<td>dct:creator example:FrankLynam .</td>
</tr>
<tr>
<td>example:ContextTypeCut dct:creator example:FrankLynam .</td>
</tr>
<tr>
<td>example:FrankLynam a foaf:Person ;</td>
</tr>
<tr>
<td>foaf:name &quot;Frank Lynam&quot; .</td>
</tr>
</tbody>
</table>

SPARQL 1.1 does not represent a major philosophical shift in the approach adopted by SPARQL 1.0 and for that reason the following explanation of the standard will be based on a reading of the SPARQL 1.1 specification. Let us now return to our Context 101 example and consider how this dataset might be queried. It has been reproduced for your convenience in Table 3.

Let us imagine that we want to find the name of the example:context101 resource. Since we have published the RDF as a text file to the http://example.org web server, we could download it from here and then write a text parsing script in Python or some other scripting language to retrieve this information. We might look for lines following example:context101 that contained rdfs:label and then extract the value from the string that followed this predicate. Very quickly, however, it becomes clear that this approach
presents a number of limitations. For example, we would need to accommodate in our script all of the possible ways in which this meaning can be validly encoded as Turtle, which are many. And then we must also download the entire RDF dataset in order to access this one value. This is fine when we are dealing with smaller files but becomes unworkable when the triplestore grows in size.

**Figure 12: the updated Semantic Web 'layer cake' (Wikipedia 2015)**

SPARQL provides a solution to this problem but some extra work needs to be done on the data provider side before it can be employed. SPARQL is a query engine that operates as a part of the Semantic Web technology stack (Figure 12). While there is no requirement to use one software or hardware solution to deliver the entirety of this Semantic Web stack, nowadays it is, nonetheless, common to employ a combination system. This fulfils a large majority of this functionality and these platforms have come to be known as RDF triplestores because most are designed to cater for RDF triple data. There are many vendors now providing RDF triplestores, such as OpenLink Virtuoso.

---

48 A digital technology stack describes the collection of digital systems used to deliver a service. Often the individuals work as a chain with the output of one system being fed as an input into the next. As such, it has become popular to describe these stacks schematically as is shown in Figure 12.
(W3C 2012), Sesame (Broekstra et al. 2002), Apache Jena (Lindörfer 2010), 4store (Harris et al. 2009) and AllegroGraph (Aasman 2006). While each presents its own particular set of advantages and disadvantages, for the purposes of this account and indeed for the remainder of this thesis, we will confine our discussion to the Apache Jena triplestore. Apache Jena was first developed by HP Labs, before being taken over in 2009 by the Apache Software Foundation. It is an attractive solution because it is lightweight, easy to install, has an API and provides support for OWL reasoning. Jena can also be installed with the Fuseki engine, which provides SPARQL access to the triples stored in the Jena triplestore. The following account assumes that the full Apache Jena Fuseki environment is installed and running correctly on a networked computer.

Now let us consider our Context 101 RDF example on the understanding that it has been loaded into a Jena triplestore. The Fuseki engine includes a web page interface that can be used to send SPARQL queries to the server (Figure 13). The client writes a SPARQL query in the text field provided, then selects a preferred format and clicks the run button to execute the query. The Fuseki engine then analyses the query criteria and returns any data that match the request in the requested format.

![Figure 13: the Apache Jena Fuseki SPARQL query interface](image)

49 DBpedia, the largest triplestore on the Semantic Web, is hosted by OpenLink Virtuoso.
50 In practice, OWL reasoning places too great a resource demand on the host server, which renders it unusable in most cases. This topic is discussed further in Chapter 3.
51 All of the RDF experiments described in this thesis were run against an Apache Jena Fuseki 2.3.0 installation running on an Ubuntu 12.04 server.
52 Loading a Jena triplestore with RDF content is covered in Chapter 5.
The SPARQL query shown in Table 4 asks a very simple question of the server. It is asking for the first 10 RDF triples in the triplestore.

```
SELECT * WHERE {
  ?s ?p ?o
} LIMIT 10
```

Table 4: a simple SPARQL SELECT query

The heart of a SPARQL query is made up of a series of tripartite criteria, which match the form of the RDF triple. Each of these triple criteria can include a subject, predicate and an object as a variable or as a constant. In the case of the SPARQL query shown above, there is only one triple criterion set and this is written on line 2 of the query. Words that are preceded by a ‘?’ sign are variables and this means that they will be populated with data by the server based on the other criteria. In this case, line 2 contains a triple criterion that contains three variables: ?s, ?p and ?o. This tells Fuseki to match any triples in the triplestore that have a subject, a predicate and an object. All triples in the triplestore will satisfy this criterion. The SELECT statement specifies the variables, which are to be returned to the user. Specifying ‘*’ here signifies that you want all variables declared in the criteria to be returned. Of course, if you submitted this query to Fuseki, it would return every triple in the triplestore, which depending on the triplestore in question, might be a lot of data. To circumvent this possibility, we can append the keyword ‘LIMIT’ to the end of the query and this will limit the amount of triples returned to the number specified, in our case ten. Running this query against the Context 101 dataset and requesting the return type to be styled XML returns the output shown in Figure 14. The table contains three columns each headed with the name of one of the query variables and the underlying cells are populated with the values for these variables.

---

53 As is the case in RDF Turtle serialisations, the use of spaces and tabs in a SPARQL query is purely aesthetic and has no functional implication.
54 Depending on the version of Fuseki that you are using, the results may be displayed differently.
Of course, asking for the first ten triples in the triplestore is not particularly informative as it could return literally anything from the resource. The next SPARQL query, which is shown in Table 5, introduces the use of constants into the query, which serve to limit the information that will be returned.

```
PREFIX example: <http://example.org/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?context WHERE {
  ?context rdf:type example:ContextTypeCut
}
```

Table 5: a SPARQL SELECT query with constants

Line 1 includes a PREFIX statement, which functions in the same way as a @prefix in Turtle. The prefixes declared can be used throughout the SPARQL query in place of the URL base that they map. The single criterion used in this query includes a single variable, ?context, and two constants, rdf:type and example:ContextTypeCut. This tells the Fuseki engine that we only want those subjects that link to the example:ContextTypeCut resource via the rdf:type predicate. To paraphrase this query: it is looking to find all contexts, which have been defined as cuts. Running this query returns the results shown in Figure 15.

```
PREFIX example: <http://example.org/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?context WHERE {
  ?context rdf:type example:ContextTypeCut
}
```

Figure 15: the results of running the second SPARQL query against the Context 101 dataset

Make sure, however, to note the slight difference in syntax between the two implementations.
SPARQL query criteria can also be chained and this is where the flexibility of the paradigm comes to fore. The query shown in Table 6 augments the previous example.\(^{56}\)

![Table 6: chaining criteria within a SPARQL SELECT query](image)

The addition of a period following a criterion adds the subsequent criterion to the list of criteria for the query. In this case, line 6 looks to retrieve all cut contexts. Line 7 then stores the resource linked to these contexts via the dct:creator predicate in the ?creator variable. ?creator is then used as the subject in the last criterion on line 8 in order to retrieve the foaf:name of that resource. Running this query returns the results shown in Figure 16 and the graph, which is crawled to construct the results, is represented in Figure 17.

![Figure 16: the results of running the third SPARQL query against the Context 101 dataset](image)

![Figure 17: the RDF graph crawled by the third SPARQL query](image)

\(^{56}\) Note that this query includes the ‘a’ keyword. ‘a’ can be used in a SPARQL query in place of ‘rdf:type’.

62
SPARQL, while challenging at first, can come to provide an amazingly flexible route into a complex RDF graph. The objective of this last section was to provide an introduction to the types of questions that can be asked using SPARQL. In practice, SPARQL queries will include far more complex functionality, explanations of which go beyond the scope of this work. While SPARQL 1.1 is the latest release of the standard, more recent non-official additions have been implemented by a number of triplestore providers, which allow for more advanced and specific functionality such as the geospatial querying afforded by the GeoSPARQL extension (Battle & Kolas 2011). And many of the RDF triplestores now also provide sets of proprietary functions, which can be used to process results before they are returned to the client (OpenLink Software Documentation Team 2015; The Apache Software Foundation 2015).

The Archaeological Semantic Web

In 1971 Robert Chenhall wrote a paper entitled The archaeological data bank: a progress report (Chenhall 1971). It outlined the design of a networked Archaeological Information System that could be used to house the data of all US-based archaeological projects. This singular interface would allow archaeologists to ask questions of the entirety of the American archaeological dataset as one unit using a computer terminal. Chenhall reasons that the best way to approach this problem is not to construct ‘one massive computer bank’ but instead to have each regional centre store and process its own data and that these individual datastores would then be linked by electrical networks. He speculates on a model for making the data available to the public, specifically whether access should be monetised or not. He strongly advocates the use of standards for the inputting of information at each regional centre, arguing that for each regional centre to develop its own competing system would be a waste of resources and also that it would create problems in respect to combining the various formats. On this last point, he notes that the format proposed in his paper has been released into the public domain, but at the same time he wonders whether it will be applicable to all forms of data that might inhabit the system. He suggests that a more flexible system should be developed ‘to be usable for all geographic areas and in the context of research problems as yet unformulated’ (Chenhall 1971 p. 169).

---

57 To learn more about constructing SPARQL queries, consult DuCharme (2013).
58 As of 1 September 2015, the linkedarc.net SPARQL endpoint has been extended to provide support for GeoSPARQL queries.
The points that Chenhall makes are remarkable for their prescience. He essentially outlines a model for the Semantic Web nearly thirty years before it was first proposed as a possibility. While it is unusual to find such a length of time separating a system’s hopeful specification and its eventual realisation, the issues that Chenhall highlights are reflective of the very nature and fundamental needs of the archaeological method as it has been constructed over one hundred and fifty years. What marks out the Chenhall story is not so much this fact but rather that the Semantic Web, when it eventually arrived, so closely matched the requirements of this existing archaeological practice. In other words, the Semantic Web was clearly not developed in response to Chenhall’s call but by chance it happened to perfectly accommodate its objectives, nonetheless.

We will now consider how Semantic Web techniques came to be adopted by members of Chenhall’s archaeological community. As with many archaeological methodological advances, the Semantic Web stimuli originated from outside the field. However, this does not mean that archaeologists and cultural historians have not played their part as innovators and contributors to the wider Semantic Web system, as we will see in the following account.

**Early investigations into archaeological digital semantics**

The history of early archaeological activity within the Semantic Web techno-social space (or the Archaeological Semantic Web, as we will refer to it henceforth) is very much tied to that of the CIDOC CRM (Isaksen 2011 p. 42). Early variants of the CRM (first a relational model and then an Object-Orientated model) predate the Semantic Web by a number of years, and so not all of these early archaeological CRM projects can be said to conform to Linked Open Data practices, but nonetheless they often espouse the tenets (linkability of data, semantic transparency) of the Semantic Web and, therefore, represent a point in a trajectory towards that Semantic Web end.

A cluster of three Greek projects marks an early waypoint for this research, which is perhaps unsurprising as the CRM has always had a strong relationship with that country (Doerr & Crofts 1999a). In 1994 the Crete-based CLIO system for the digital management and storage of cultural heritage artefact descriptions explicitly called for the need to create a semantic model that allows for the ‘dense linking of information, access by unlimited chained references, expression of abstract properties and various ways of joint temporal and spatial assignment’ (Constantopoulos 1994). These were all
aspirations largely consistent with the later Semantic Web. A year later, the POLEMON project continued this thread by stating its objective to support the work of the Greek Archaeological Service by developing a ‘computerization of the National Archive of Monuments’ in four locations in Greece and also to create a ‘full-scale museum information system in the Benaki Museum, consisting of two subsystems (collection management and cultural documentation)’ (Pantos & Bekiari 1995 p. 3). In 1999 Doerr and Bekiari, who had worked on the POLEMON project and who would be instrumental in the formulation of the CRM, talked about the challenges of representing the semantics of partial artefacts within a conceptual model (Bekiari & Doerr 1999). These three projects drew from the earlier development of the CANDIA system by FORTH/ICS, which was released in 1989 and provided ‘read, write update and printing facilities’ for Greece’s archaeological monuments’ records (Pantos & Bekiari 1995 p. 2) and, in many respects, marked a first significant step on this road towards the digital Archaeological Information System. The defining characteristics of this period are a focus on the production side of the digital knowledge equation, on the involvement of trained informatics engineers in the building of these systems for an archaeological ‘client’ and on the reliance on the relational database model.

While Greece was important in this early phase of the Archaeological Information System, it was not the only region to experiment with new approaches to the problem of managing, storing and representing archaeological knowledge using digital means. For instance, Hansen (1999) discusses the ambitious plans of the Danes to link their existing archaeological data repositories in 1999 and to make them available via the web. While he talks about employing conceptual modelling techniques in the process, it is not clear whether or not they explicitly used the CRM as their base. The resources mentioned were also beginning to become a lot larger. Hansen names two systems that are involved in the Danish amalgamation. The first, DKC, held geotagged archaeological data from 140,000 locations and GENREG, a repository for related imagery contained about 200,000 records.

Britain, as one of the founding fathers of the modern archaeological movement has a strong history of involvement with Archaeological Information Systems. In the digital era, a defining moment in the British context happened in 1996 with the founding of the Archaeological Data Service (The University of York 2015). While the initial proposal
described a network of installations located across Britain, the ADS soon came to be primarily associated with the site and staff of the University of York. One of the key roles of the ADS over the years has been to engage with the various cultural heritage bodies in England, Scotland and Wales bringing together decades of experience of handling material assemblages and all the sub-systems of activity (cataloguing, typologising, conserving, et cetera) that accompany this type of work. In 1997 the ADS organised a Resource Discovery Workshop for practitioners, which focussed on the practical matter of introducing semantic recording into digital field practice (Miller & Wise 1997). The practical instructions outlined at this event centred on the implementation of the Dublin Core model.

In 1999 Quine presented a CAA paper discussing how the Royal Commission on the Historical Monuments of England (RCHME) was making strides towards standardising its digital data practice (Quine 1999). The MIDAS system is described as a metadata standard in this respect and it would prove to be an influential model for all UK work of this type in the years to come. However, there is little said in Quine’s piece about how all this information was to be linked and disseminated. In 2000 a weighty contribution to the field was made by Brown and Perrin, who outline the plans for the newly established English Heritage Centre for Archaeology (Brown & Perrin 2000). An aspect of the centre’s plans was to establish a new conceptual reference model to manage the data being created by English Heritage archaeological partners. The CAS Projects and ARCHives model (CASPAR) was the result of this process and this was based on aspects of the CRM and in-house conceptual models, which had been previously used by the RCHME. The report notes that these previous standards were ‘primarily concerned with monuments and fieldwork’ data and, therefore, that the role of CASPAR was to cover an existing inability to adequately model archaeological archive information.

**A new millennium, a new paradigm**

With the advent of the new millennium, we begin to see archaeological projects adopting CRM-based systems with increasing regularity. Still, however, these implementations could not be called Semantic Web resources, in the sense that we understand them today. For example, the 3D MURALE project presented in 2001 (Cosmas et al. 2001) was a tool for recording, reconstructing, encoding, visualising and querying of archaeological data, which placed an emphasis on its visual outputs, while making sure to underline the importance of having a solid semantic base in the form of the CRM. There is still no
mention of URIs, however, or of using the web as something more than a storefront for a closed dataset. The 3D MURALE project is interesting in other ways. For the first time we get a sense of the sort of technological stack employed to deliver these sorts of system and the talk of XML content being styled using XSL and 3D content being delivered via the VRML standard all reference the evolving Web 2.0 stack. These would also influence the technological course of the Semantic Web.

The first sign of a next step in the evolution of the basic paradigm employed by these Archaeological Information Systems appears in a paper given by de Haan in 2003. In Design for Global Access to Cultural Heritage, de Haan specifies the need to deliver hardware, software and semantic interoperability if cultural heritage content providers are going to allow ‘end-users of different backgrounds with uniform and universal access to a variety of cultural heritage information through an adaptibility [sic] and adaptivity using a system aiming to provide contextualised information regarding different perspectives, levels and dimensions’ (de Haan 2003 p. 1). The I-Mass project, which he introduces, is really quite forward thinking in that its objective is to create what he calls ‘Virtual Reference Rooms’ that will allow users to access cultural content from various sources. I-Mass employed the CRM as its underlying model and, critically, de Haan also mentions the importance of using URIs for ‘resource description’ and RDF and Dublin Core for ‘registry or ‘general’ purposes’. De Haan’s English is a bit unclear but what is certain is that the principle of using the infrastructure of the web to facilitate the amalgamation of multi-sourced heterogeneous data using URIs to identify data resources and RDF as a Knowledge Representation is a fundamental part of the approach taken by the I-Mass project and, as such, this marks it out as a key milestone in the history of the subject.

It would take just one more year, or three years after the publication of Berners-Lee et al’s Semantic Web, for the first contribution to appear, which unequivocally addressed an archaeological subject matter using a Semantic Web approach. Doerr et al.’s (2004) address to the attendees of the CAA 2004 conference in Budapest spoke of how a CRM-structured dataset being presented using an RDF representational system could be used to bring together ‘scientific databases and corpora describing finds with inscriptions and iconography of the Roman era’ (Doerr et al. 2004 p. 1). The problem highlighted by the

---

59 Interestingly, despite the novelty of the approach proposed, Doerr et al. do not reference Berners-Lee or any of his publications in their paper.
paper was of ‘partly complementary, partly overlapping and partly contradictory’ (Doerr et al. 2004 p. 1) archaeological information and the solution was to employ these novel Semantic Web approaches to bring together but not homogenize the disparate sources. The practical work described by the paper was primarily concerned with the converting of existing archaeological digital information found within a series of relational databases into one RDF graph whole. This demanded that classes of information (or entities in CRM-speak) be extracted and then mapped onto CRM classes and selected typological thesauri. In one strike, Doerr et al. had set a template for archaeological Semantic Web research for the coming decade and for which this current work can claim not a small amount of lineage.

From this point on, the Computer Applications and Quantitative Methods in Archaeology (CAA) conference becomes the primary platform for discussions of Semantic Web practice within Archaeological Information System contexts. As before, the CRM dominates the narratives of the CAA papers that refer to modelling practice (Andreussi & Felicetti 2008; D’Andrea 2011) but it would be wrong to suggest that the 2000s did not propose any alternatives to CRM orthodoxy. Zhang et al. (2002) reject its use in favour of another conceptual model of their own making for the modelling of generic archaeological data. Similarly, Atarashi et al. (2000) forego the CRM, preferring instead to use the established Dublin Core ontology to structure the metadata for their collection of photographs of Japanese archaeological material. There is also the example of the ArchaeoML standard, which David Schloen presented in place of the absent Kansa (2006) at CAA2006 and which was used to successfully structure the Open Context project for many years (E. C. Kansa et al. 2010 pp. 10–14).

**Bringing archaeological concepts into the CRM fold**

The 1990s and 2000s had seen an increased level of professionalism with the growth of commercial archaeological units in the UK, which were needed to accompany the huge growth in house building that occurred during that time (Aitchison 2009). This brought with it the introduction of digital archaeological recording practices that were more holistic in nature (Richards 2008 p. 174). Gone were the days when an archaeological project’s digital correlate was focussed entirely on its artifactual outputs. The plummeting costs and lower training demands for introducing digital technology into field archaeology had pushed its use closer and closer to the trowel’s edge. As a result, archaeologists of the new millennium wanted to be able to digitally model and record the
contexts that they were excavating, the methodologies that they were using and the increasingly specialised viewpoints (archaeobotanical, osteoarchaeological, palaeoclimatological, et cetera) that they were employing. In many cases, they were already recording this information digitally using their own database schemas implemented using off-the-shelf database and spreadsheet software. In that context, the only way that the openness and interoperability promised by the Semantic Web could be sold to this new cohort of field practitioners was if, and when, the Archaeological Semantic Web’s models caught up with the recording practice on the ground.

Despite the CRM’s successes within the community, with time even its strongest proponents had to admit that its capacity to fully realise its potential in the sector would always be limited by the fact that it had not been designed with the needs of archaeologists primarily in mind. To address this perceived lacuna in the CRM, a group of archaeologists and informatics engineers began working on a series of projects that would eventually lead to the drafting of a CRM extension, which allowed for a better accommodation of archaeological information (Binding et al. 2008; Cripps et al. 2004; Cripps & May 2004; May et al. 2011; Tudhope et al. 2011). This extension, which was eventually commissioned by English Heritage and, therefore, named the CRM-EH, was designed to model information derived from single context archaeological excavation. However, despite the fact that the new extension was meeting a real need within the community, the CRM-EH failed to be taken up to any great extent by Archaeological Semantic Web applications.

Despite the failure of the CRM-EH to ignite the hearts and minds of Archaeological Information System designers, this did not signal the end of efforts to create a discipline-wide conceptual model for the Archaeological Semantic Web. In 2013 the Advanced Research Infrastructure for Archaeological Dataset Networking in Europe (ARIADNE) project was created in an effort to bring together all of the disparate datasets (Semantic Web, current and future) that make up the European archaeological data scene (ARIADNE 2015a). In many ways, this can be seen in the context of a wider trend within the European cultural heritage research sector towards amalgamation and the creation of interstate networks of research (Haslhofer & Isaac 2011; Purday 2009).
ARIADNE is currently in its market research phase. This means that it is collecting data from the key archaeological stakeholders in Europe, specifically by finding out about their intellectual and practical expectations of a new shared archaeological data infrastructure (ARIADNE 2014b, 2015b). It is also conducting its own research into related subjects such as data sharing (ARIADNE 2014a) and on how it might go about supporting existing research paradigms like data mining (ARIADNE 2015c) and Natural Language Processing (ARIADNE 2015d). In the context of this history of the Archaeological Semantic Web, the Ariadne project has also been engaged with the building of a new ontology to be used for the modelling of the archaeological data that will come from its European partners and which it will eventually aggregate. The first publically released version of the CRMarchaeo, as it is named, was released in early 2014 and a version 1.2.1 update appeared in October 2014 (Cripps et al. 2014). Like the CRM-EH, it is designed as an extension of the CRM. At this point, it is too early to gauge the impact that the CRMarchaeo will have on the various different archaeological sectors that it targets but it is surrounded by a much greater public relations machine than accompanied the CRM-EH and the early signs are that it is being relatively well received.

Conclusions

Whereas Chapter 1 posed the central question of this thesis, this chapter began the process of answering it. If anything, the historical narrative recounted in these pages has shown that the problem of heterogeneous and incompatible data on the internet is not unique to archaeology. Berners-Lee and his colleagues proposed the model of the Semantic Web because they felt that data had been largely forgotten about in the web’s evolution. In their view, in order for the web to reach its fullest potential, data needs to take centre stage in the next generation web. However, despite an initial flurry of interest in the idea, it would be difficult to argue that the Semantic Web, whether in its broadest sense or in the narrower context of its role as a platform for archaeological research, has been a resounding success to date. In Gartner Hype Cycle terms (Linden & Fenn 2003), we might say that the model has passed its ‘peak of inflated expectations’ stage (this happened early on, ending in about 2002) and while there are signs that it is now onto the ‘slope of enlightenment’, it could also be argued that it may be still languishing in a ‘trough of disillusionment’ (see how these phases are reflected in a chart showing the quantity of publications on the subject by year, represented in Figure 18). One possible reason for this lack of take-off could be the model’s perceived complexity, which is a view that comes up repeatedly in its discourse. For whatever reason, Semantic Web
technologies have been labelled a complex challenge for which the benefits are not clearly seen to outweigh the costs for the data producer (Isaksen et al. 2010 p. 1). If this were indeed the case, then it might go some way towards explaining another trend that emerges out of this chapter, which is that the archaeological projects that have most fully immersed themselves in the Semantic Web ethos and practice are almost exclusively associated with the larger archaeological and cultural heritage institutions. The increased financial and technical resources at the disposal of these groups have allowed them to bear the considerable challenges that come with Semantic Web implementation. For now though, let us hold our final council on this matter of difficulty, at least until we have considered the building of our own RDF dataset in Chapters 4 and 5.

![Figure 18: a histogram showing the dates of Linked Open Data-related resources selected by LODLAM](image)

What is indisputable, though, about the Semantic Web model is that it fits extremely well with existing archaeological practice. The fact that all Semantic Web data asserts its meaning through its relationships with other data is essentially an archaeological view of the world and of knowledge. As such, there is no need to give archaeologists the ‘hard sell’ as to the merits of the Semantic Web’s epistemology; this way of thinking is already built into their DNA. There is also the matter that the Semantic Web is now a relatively

---

60 Linked Open Data in Libraries, Archives and Museums or LODLAM is a primarily web-based group who promote the exploitation of Linked Open Data resources in the galleries, libraries, archives and museums sector (often referred to by the acronym GLAM). LODLAM curates a Zotero bibliography, which contains over 450 entries that have been contributed by members selected on the basis of their relevance to the field. This histogram shows the publication dates and/or launch dates for these books, journal articles and GLAM Linked Open Data projects.
mature technology – at nearly two decades of age in 2015. As a result, it has gathered about it a critical mass of technical experience and archaeologists have been involved in their own Semantic Web experimentation for quite some time now. These have highlighted the approaches that work and those, which do not, and because of the transparent nature of Semantic Web research, these insights have become part of public knowledge. This is also a valuable asset of the Semantic Web that needs to be borne in mind.

These conclusions, however, tell us only half the story of what we need to know about the Semantic Web model before we can continue in our narrative. We have yet to address the feature that gives the Semantic Web its ability to act as a semantic medium between data producer and consumer. As such, in the next chapter, we turn to the subject of ontologies and ask how these data structuring devices inject meaning into Semantic Web content.
Chapter 3 – Data transparency and re-use

An examination of the ontology in its philosophical and Knowledge Representation senses

‘The essential nature of the propositional sign becomes very clear when we imagine it made up of spatial objects (such as tables, chairs, books) instead of written signs.’ (Wittgenstein 1922 sec. 3.1431)

Introduction

In Chapter 2 we presented the Semantic Web as a solution to the problem of heterogeneous data within archaeology. Using RDF, we showed that it was possible to produce archaeological content in a form that would allow it to be embedded within an existing mesh of domain knowledge. The argument being that this linked nature of the data increases its inherent value and as a consequence the value of the wider pool of the knowledge. SPARQL was presented as a route into what can appear on first inspection to be a largely unintelligible mass of information. It simplifies the querying process, attenuating its complexity and allowing individual nodes of information to be plucked from vast graphs of networked knowledge. We also made the apparently logical observation that data that is meaningful is more valuable to the research community than data that is absent of meaning, and RDF and SPARQL need to be understood as avenues into this meaning. However, we have yet to consider the nature of this meaning and how this meaning might be injected into an RDF dataset and subsequently extracted by a machine agent. The answer to this question is found in the form of what is known as a Knowledge Representation (KR) ontology and this chapter will show how essential the ontology has become to the practical workings of the Semantic Web.

To view the Semantic Web ontology simply as a practical solution to a functional challenge would, however, be a mistake. The study of ontologies is a central question in philosophy with its origins being traced back to the beginnings of when humans started to ask deeper questions about themselves and about the nature of their relationship with the world around them. Therefore, to say that a Semantic Web ontology gives meaning to an RDF dataset and that this satisfies the ontological matter would be to entirely
misrepresent exactly what the Semantic Web is attempting to achieve; that is to represent views onto aspects of reality, whatever those interpretations might be.

To give some context to this understanding of the ontology, this chapter will begin by asking what an ontology is in its broadest sense. This history will begin at the advent of the Western philosophical tradition in the teachings of the Greek Presocratic philosophers and end with the ontology’s employment as a practical tool of KR scientists in the modern era. We will then look at some of the most important ontologies currently in use within the Archaeological Semantic Web. The CIDOC CRM and its CRM-EH and CRMarchaeo extensions, whose historical trajectories were introduced in the last chapter, are deconstructed and explained from the perspective of both the Semantic Web data publisher and consumer.

**Why the philosophical perspective?**

The *OED* offers the following definition for the term ‘ontology’. It is a ‘system similar in scope to modern predicate logic, which attempts to interpret quantifiers without assuming that anything exists beyond written expressions’. This understanding of ontology would seem to fit well with this thesis’s focus. It sees the ontology as a tool used within the study and application of logic and the Semantic Web is, if nothing else, concerned with the formulation of logical rules and the outcomes of their application. The *OED*, however, also offers a second definition. In this case, the ontology is described in its philosophical sense. It states that ontology is ‘the science or study of being; that branch of metaphysics concerned with the nature or essence of being or existence’. On first inspection, this presents quite a different understanding to the previous statement. If we want to approach a fuller understanding of the ontology in the context of its use within the Semantic Web model, is it necessary that we also attempt to understand what on the face of it is a far more complex and broad philosophical range of issues? And given that this is not a philosophical treatise, is this a valid use of our time? After all, the majority of works that deal with the Semantic Web do so from a primarily, if not exclusively, functionalist point of view.

While valid arguments can be made to the contrary, I would contend that it is worthwhile to pursue this generally neglected subject, albeit in a necessarily concise fashion. As was stated in the introductory chapter, a primary objective of this thesis is that it be written from a reflexive point of view, insofar as is possible. The first chapter also expressed the
Chapter 3 - Data transparency and re-use

hope that the reader of this work be open to this particular way of engaging with the material. In that context, it would be disingenuous to now ignore this aspect of the subject, particularly given that it will prove fundamental to our later study of the wider implications of the Archaeological Semantic Web.

Where to begin?

Karl Popper said that there are two fundamental questions, which a person can ask. The first is, what is the nature of the world around us and the second, how can we obtain knowledge about that world (Popper 2012 p. 8)? And so, in Popper’s eyes, philosophy can be broken down into two basic categories of enquiry, which we call ontology and epistemology. It is not difficult to see that there is a connection and dependence between these two lenses of enquiry. Epistemology, for instance, cannot exist without an object to study and the abstract sense of these entities is agreed upon by an ontology. Therefore, we can say that epistemology relies upon there being a particular model of being on which to operate and in that sense ontology precedes epistemology. However, in practice the two are tightly bound together and it is difficult to consider one in isolation of the other. As such, this following account may include passages that address epistemological as well as ontological concerns.

If we are to tackle this subject of the philosophical ontology, then we need a starting point. Unfortunately, but perhaps predictably, there is no neat point in time, which can be found to satisfy this need. It is highly likely that humans have been preoccupied with matters of their own existence and their place within the wider world for as long as they have had the mental capacity to do so. As a result, the true chronology of ontological thought is undoubtedly a very deep one (Lewis-Williams 2002). However, we will resist the temptation of delving down that particular rabbit hole and instead begin our journey on the steadier intellectual footing of the first Western philosophical thinkers, in the late 6th and 5th Centuries BCE philosophy of Magna Graecia.

Thales is widely acknowledged to be the first philosopher to espouse a set of principles, which we now define as Presocratic (Mansfeld 1985). And, as is the case for most of this collection of early scholars, what we know of his work comes to us indirectly through the writings of other philosophers, most notably the 4th Century BCE Aristotle and his student, Theophrastus (McDiarmid 1953 p. 88). While not ideal, and while we need to be aware of the distinct possibility that many of these accounts will be effected by more
than a small degree of contamination, they are all the evidence that we have in many cases and, therefore, we must make the best of them (Cherniss 1951 p. 319; Waterfield 2000 pp. xiii–xiv). Thales’s cosmology paints a picture of an earth resting upon a vast body of water, as a ship floats upon the sea (DK 11A14-5).61 Continuing this aquatic theme, Thales also believed that all matter comes from water (DK 11A12) and because of this Aristotle felt that Thales anticipated his own material cause of being axiom; this stated that all entities had a material cause (Lloyd 2012 p. 1). Thales also believed that the being was divided between the soul on the one hand and the external material form on the other (DK 11A22).

Anaximander, a fellow Milesian and a near contemporary of Thales, argued for a different reading of the universe’s form, although Popper argues that his is likely to be a critical reaction against the earlier philosopher’s model of existence (Popper 2012 p. 10). Anaximander’s worldview is of an earth that is shaped like a drum (DK 12A10) with two opposite-facing surfaces. This hypothesis appears to anticipate the science of gravity and relativity when he states that ‘we walk on one of its flat surfaces, while the other is on the opposite side’ (DK 12A145). Anaximander and the other Presocratics lived in a time of great public and private transformation in the Greek world (Osborne 2007). While we need to be on our guard against overstating the effect of these changes, it is clear that the great minds of the Greek city states were struggling with the challenge of charting a course from a point in which knowledge was either divine and unobtainable or mortal and knowable to a point where knowledge creation was firmly in the hands of humans or as Finley puts it, ‘it is as if the early mist had risen to uncover no longer a world of gods but a bright mid-morning’ (Finley 1966 p. 80). Anaximander’s great contribution was to begin a process of rationalising the nature of the relationship between divine and mortal beings on the one hand and divine and mortal knowledge on the other. And he attempted to do so by proposing hypotheses that might explain these concepts within the context of a fairly rigid set of permissible parameters. His ‘apeiron’ concept can be seen in this context (Waterfield 2000 p. 5). It is an infinite and intransient entity from which all being emerges and to which all being returns (DK12A9). Anaximander expands upon the singularity of Thales’s water model by stating that the apeiron is also the source

61 The DK abbreviation refers to Diels and Kranz (1951). This is a compendium of many of the often extremely limited remains of pre-Socratic philosophical teachings.
of opposites (cold and hot, rough and smooth, et cetera) and that these introduce variety into beings.

A key figure in this period is the Elean, Parmenides, whose ontological proposition bucks the general trend towards seeing change as a fundamental aspect of reality.\(^6\) The only surviving work accredited to him is a type of sacred poem, which is now separated by experts in the field into three units: a prologue, a section called ‘The Way to Truth’ and a final section entitled, ‘The Way of Appearance’ (Waterfield 2000 p. 49). The concept of the poem is a journey taken by the narrator from the land of mortals down into the underworld.\(^6\) While the events of the piece are structured around this physical expedition, the real journey is a metaphorical one from a position of ignorance to an understanding of how things really are. This knowledge is revealed to the self-proclaimed mystic, Parmenides, by an unnamed goddess, who offers Parmenides a choice of two paths to the way of truth, but the inference is that only one of these options holds the possibility of obtaining that goal (DK 28B2). Presumably, this implies that humans have up until now erred in their appreciation of exactly what and how the world is. To say that Parmenides’s poem veers on the side of opaqueness would be an understatement and as a result the intended meaning encoded within its narrative has attracted vigorous comment and counter-comment over the years.

While it would be convenient in the context of this study of ontology to see ‘The Way to Truth’ as a treatise on Parmenides’ position on ontology and ‘The Way of Appearance’ as the poet’s views on epistemology, the reality is that there is no such clear distinction between the two topics and in that sense we must be careful when reading the first half of the poem as an unconditional ontological work. Having said that, we can definitely use the poem in its entirety to attempt an outline in broad terms a Parmenidean ontological model. The goddess’s role is to guide Parmenides in what it means to say that something ‘is’ (DK 28B6). She explains that there is a difference between thinking that something exists and for that something to actually exist (DK 28B3). There is also the problem of what is not. This conceptualisation of negation or absence has troubled philosophers

---

\(^6\) Popper notes how the period can be broadly defined by a movement away from an ontological model which saw stasis as the norm, as is testified in the writings of Homer, to a cosmology of constant change by the time of Aristotle (Popper 2012 p. 14).

\(^6\) See Kingsley for an alternative shamanic take on the meaning of Parmenides’s journey (Kingsley 1999).
ever since (Azzouni 2012; Crane 2012) and, while the intent of the passages that deal with this subject are far from straightforward, the contribution of Parmenides is mainly that he is voicing these concerns for the first time. The poem also alludes to the fact that in order for an entity to exist, it must have attributes associated with it. Parmenides also learns that all entities are immune to death and are unchanging both in terms of space and time, which is a critical insight given the overall context of ontological stasis (Waterfield 2000 p. 53).

The Pythagoreans, an ascetic group of philosophers, who were almost cult-like in their devotion to a particular way of thinking and way of life, are named after their founder Pythagoras, a Greek Samian, who emigrated to the south coast of Italy, possibly fleeing the persecution of the tyrant, Polycrates (Waterfield 2000 p. 89). Unfortunately, very little of the work of Pythagoras or his followers has survived and we are left with the familiar problem of unpicking these original beliefs from the tangle of later references, in this case mainly Plato’s (Riedweg 2005 pp. 116–118). However, it does appear to be the case that Pythagoras held an ontological viewpoint that diverged markedly from the materialist view espoused by the Milesians. The Pythagoreans held that the ultimate reduction of all being was numeric (Aristotle fr.203 Rose). In that sense, the characteristics that applied to numbers – being even or odd, their limits and so on – could all be applied to things, as well. Aristotle suggests that limits for the Pythagoreans are, therefore, in some way fundamental to the nature of a thing (Waterfield 2000 p. 91). Unsurprisingly, mathematics in the form of arithmetic is a key analytical tool for the Pythagoreans. It is not simply used to solve abstract problems but is instead a means by which the true reality of being might be uncovered (Nidditch 1983 p. 12). In this light, the Pythagoreans advocated a cosmology known as the ‘The Harmony of the Spheres’, which relates the universe’s structure to the mathematical rules governing sound and music and which proved immensely popular as a model for later cosmologists (Kahn 2001 p. 166).

**Platonic and Aristotelian ontological models**

By the time of Plato and his pupil, Aristotle’s arrival on the scene, a number of quite different models for how the world is and how it reveals itself had already been proposed. The era of the Presocratics had been characterised not by a dogmatic uniformity of opinion often associated with the idea of a school of thought in the modern sense but by a willingness to critique the established position (Popper 2012 p.
10. Socrates, Plato and Aristotle continued this tradition as they proposed their own ontological models and it is within their work that we can begin to clearly identify the basis of an ontology that is recognisable in the modern KR sense.

Plato’s primary contribution to the ontological debate was his ‘Theory of Forms’ (Dancy 2001). The Theory of Forms or ‘Theory of Two Worlds’, as it is also labelled, defines a bipartite model of reality and this is introduced most famously in the philosopher’s Allegory of the Cave (Fogelin 1971). On the one hand, there is the real world that is occupied by entities known as forms and then there is the world of illusion, which we, as humans, inhabit. Plato’s ontological model is very much predicated on the idea that the real forms cannot be accessed or known directly (Cooper 1970). Only representations of them, which are illusionary and, ultimately not truth, can be accessed. In this sense, Plato’s ontology presents a fundamental epistemological problem that apparently cannot be overcome. Theoretically, it is entirely independent of knowledge but it can only be understood as an object of epistemological agency (Calvert 1970 p. 41).

Plato’s forms and the illusionary ‘particulars’ that we perceive are related and there is interplay between the two. It is tempting to view the Platonic form as an ideal prototype for its illusionary counterpart. While the forms are universal and never come into or out of being, the particulars change, are born and decay and go out of existence. Lacewing (2007) illustrates this point by considering an example where an object shows reflections of beauty. The object may die and, therefore, go out of existence but Plato would argue that the fact that the beauty form does not also die and that other particulars can be seen to reflect this form henceforth, shows that the forms and their reflections are not tied to one another.

Aristotle, for his part, argues against this two-world reality dichotomy (Fine 1993). For Aristotle, there is only one world of reality and that is the natural world (Lavine 1985 pp. 70–71). He attacks the Platonic Theory of Forms on the following basis. First, he turns on its head the Platonic idea that knowledge comes from the unchanging real world of the forms. For Aristotle, the unchanging world is the natural world, which is made up of what he calls substances and these are the source of knowledge. He applies his Aristotelian logic to find that Platonic forms must be merely copies of real concrete
things and that the two-world model is unsustainable because it makes it impossible to explain how the sensible and intelligible things might be related.

The ontology of Aristotle is, like Plato’s, tightly bound to his epistemology. To know, for Aristotle, is to understand the causes of a substance and in his ‘Four Causes’ theory, the Aristotelian ontology is defined as a function of its unveiling via his epistemological method (Falcon 2015). First, Aristotle tells us that we must look to the material cause of a thing. By applying this method, we learn about the materiality of a thing; for example, whether it is made out of wood or metal. Following the same logic, its formal cause tells us about its shape, its efficient cause tells us how it was produced and, lastly, its final cause tells its function. Aristotle does not believe that there is a separation between the things of perception and those of truth. His is an ontology in which form and matter are delivered in the one package, as an ontological unity (D. Charles 2001). He does not believe that perception is to be mistrusted as Plato does and, as such, Aristotle is the prototypical empiricist and positivist. In that sense, he is the father of the inductive method, which would prove so influential in the development of modern science (Popper 2012 p. 6).

The paucity of sources that survive for the Presocratics makes it difficult to clearly define the ontological views of these early philosophers. The situation improves dramatically by the time of Plato and particularly Aristotle as both go to some lengths towards explicitly outlining their thinking on the ontological question. It is in the writings of Aristotle that we can begin to separate out the field of metaphysics under two discreet question headings (Hofweber 2014). What is there? And what is it like? In that sense, ontology becomes a sub-field of metaphysics, that being the study of the basic categories of things and this specialisation inevitably affirms a strong link between ontology and the form in both its Platonic and Aristotelian understandings from this point on.

---

64 Despite the fact that the term ‘metaphysics’ is overwhelming attributed to Aristotle, the philosopher never actually used the term to refer to any of his works (S. M. Cohen 2014). The term literally means ‘after the physics’ and comes from a later arrangement of Aristotle’s writings, probably at some point during the first century CE. The editor of the later collection placed the writings that dealt with what Aristotle himself called the First Science, Wisdom or Theology (Collingwood & Martin 2001 pp. 5–6), after a section dealing with Aristotelian views on the subject of physics; hence the name.
The European Enlightenment

In the ancient world, what we now call science was known as natural philosophy, and much of the method used in this study was set out by Aristotle in his collection of treatises (Heidegger 1991 pp. 62–63). For centuries, very little changed in the manner of the natural philosopher’s engagement with their subject matter. A lack of any major technological advancement played a part in this tendency towards inertia but it was not the only factor; the church and religion in general also had vested interests in maintaining the status quo, and anything that challenged their foundational teachings was naturally mistrusted and discouraged (Grant 1971 p. 4). Even by the arrival of the Middle Ages, nearly a thousand years after many of these ideas were first postulated, scholars still used the writings of ancient Greek and Roman philosophers as the basis for many of their views on a wide variety of subjects: mapping and cosmology, politics, ethics and natural philosophy (Highet 1949 pp. 1–2). In respect to ontological viewpoints, in Europe scholars, who were often men of the cloth, had managed to accommodate the Aristotelian framework within an overarching Christian paradigm, and none made a greater contribution to linking the two traditions perhaps than the Neoplatonist, Augustine of Hippo (O’Meara 1981 p. 39).

A major epistemological shift against this intellectual stasis was instigated by the coming of the European Enlightenment in the 17th and 18th centuries (Cassirer 1951 p. 4; Pagden 2013 pp. 1–18). Of course, the Age of Enlightenment did not emerge out of thin air and much of the contributions that would later define its output – the primacy of reason, humanism, social and political change, technological innovation – had their roots in the Renaissance (Kelley & Popkin 2012). By this stage, the idea that method was a vital cog in the study of natural science was well established. Figures such as Galileo, his contemporary Francis Bacon and then Newton all recognised and illustrated the greater insights provided by sound methodological and increasingly sophisticated analytical frameworks.

By the 19th century, science had finally severed its link with its progenitor, philosophy, when the experimental scientific method, first implemented by figures such as Newton, Boyle and Locke, had become well established (Ben-Chaim 2004). A number of philosophers were also keen to profit from the methodological advances being made in the sciences by adopting and adapting these more explicitly scientific approaches to the
study of subjects traditionally associated with philosophy. This movement came to be known as analytical philosophy (Longworth 2009). The approach of the analytical philosophers gradually came to dominate the field as it moved into the 20th century. In fact, so influential has the analytical approach proved that now the philosophical discipline can be broadly categorised into an Anglo-American school on the one hand, which continues to practice the methods established by the first analytical philosophers and the so-called continental philosophers of mainland Europe, who do not (Scruton 2012).

An analytical philosopher is marked out first and foremost by their focus on method. And the objective of this method is primarily about breaking down a problem into its component parts. For example, an analytical philosopher would consider the problem of the nature of existence by first establishing a number of propositions that they believe to relate to the matter. They might in the first instance state that all humans have the possibility of existing. And that existence begins with a human birth and ends with a human death. These are testable statements, which can ultimately be verified or discounted by being subjected to rigorous logical analysis using a more scientific type of language that we call formal logic (Martinich 2008 p. 2). Disproving any one of the propositions serves to discount the associated model. If this happens, the model must be tweaked and the propositions logically retested until all are proved true.

Gottlob Frege was the first to suggest that philosophy needed a new grammar with which to provide solutions to its many remaining questions. He devoted his career to the study of a single conundrum, that of the logic of arithmetic, and he did this using a new formal language, which he invented to test out his model (Dummett 2008 p. 7). While Frege used his formal language to test out his theories about the nature of arithmetic, he endeavoured to ensure that the language was subject-neutral and that it could be used to logically interrogate any problem (Dummett 2008 p. 9).

It is debatable whether Frege, were he to be alive today, would consider himself to be an analytical philosopher but it is likely that G. E. Moore would. Moore’s work can be seen as a reaction against the grand theories of the German Idealists (Baldwin 2010). He saw this approach to philosophy as being a false science, relying on the type of transcendental enlightenment employed by figures such as Parmenides. Moore wanted philosophy to be
taken seriously by the academic community and this meant employing testable and demonstrable methods. Moore (1899 p. 182) stated that ‘a thing becomes intelligible first when it is analysed into its constituent concepts’. The analytical philosophers wanted to start with the empirical data and to use this to induce a theory that would satisfy the reality of the data.

Bertrand Russell is considered by many to be the most influential of the analytical philosophers (Beaney & Reck 2005 p. 11). By applying techniques honed in mathematics, Russell produced a quantificational logic that he used to deconstruct language propositions, as had Frege and Moore before him. The objective of Russell’s linguistic approach was to get at the logical form of any statement and to then use that form to suggest answers to philosophical questions. In this sense, Russell was an atomist (Irvine 2015). He believed that any proposition could be broken down into a series of atomic truths and that these truths would then reflect an objective reality.

Ludwig Wittgenstein was Russell’s protégé, and at first the Austrian continued to develop the thread of his master’s teachings (Proops 2013). While Russell was interested in applying formal methods to answering all philosophical questions, Wittgenstein came to see language as the gatekeeper of all philosophical truth (Biletzki & Matar 2014). In his *Tractatus Logico-Philosophicus* (1922), he attempted to reconcile Russell’s atomism with Frege’s apriorism as an approach to this problem of reducing language to truth. The *Tractatus* concludes by stating that there are two types of proposition. The first have as their subject the physical world and can be reasoned to be either true or false. He presents his so-called ‘Picture Theory’ to explain this first type (Pears 1977 p. 183). The truth of a statement can be arrived at if the subject of the statement can be visualised. This idea is what is now known as a correspondence theory of truth (Kirkham 1992 pp. 119–140) and can be related to the Platonic theory of forms. The second type of Wittgensteinian proposition does not represent a subject matter of the real world, and, as a result, its truthfulness cannot be assessed.

Wittgenstein would later go on to deny these conclusions and although he never officially published any other treatise on the matter, his Cambridge lectures on the subject have appeared posthumously in a collective work called the *Philosophical Investigations* (1973). In this, he maintains the position that language is split into two types of statement, personal
and physical, but he denies that philosophy can answer any of the core questions, including that of ontology. By the end of his life, philosophy represented for him a form of therapy that could only be used to help one deal with the fact that truth was ultimately unobtainable (McGinn 2013 p. 23).

Despite Wittgenstein’s later rejection of the possibility of breaking down language into its component parts, which could then be judged true or false, the general tenets championed by the analytical philosophers did not end with the work of Frege, Moore, Russell and Wittgenstein. In fact, the approach proved to be hugely popular and influential not just in philosophical circles but also as a practical methodology in answering related questions within other disciplines. One group that took particular inspiration from the analytical model was a philosophical reading circle, which gathered every Thursday around the figure of Moritz Schlick in the Austrian capital of Vienna between 1924 and 1936 (Uebel 2014). For the most part, the Vienna Circle’s approach to answering philosophical questions was entirely consistent with that of analytical philosophy. They rejected German Idealism and embraced the testable propositions created by the symbolic logic languages proposed by their predecessors. However, while Russell was primarily interested in the logic of mathematics and Wittgenstein the logic of everyday language, the Vienna Circle was far more expansive in the subjects that they discussed and considered relevant to their way of thinking. For instance, they felt that logical positivism, as their movement came to be known, could contribute to the arts and architecture (Galison 1990), economic and social planning (Pietarinen 2010 p. 73) and politics, with all their members holding entrenched socialist views (Uebel 2004 p. 250). Because of their adherence to the ‘data first’ doctrine, they came to propose several controversial positions on many longstanding philosophical subjects, such as ethics and morality. Carnap (1935) and Neurath (1983), for instance felt that these were representative of shared human opinions and because they could never be proved true or false, this rendered them ultimately meaningless. This rejection of any form of transcendental influence on philosophy culminated in a gradual sidelining of metaphysical questioning (Feibleman 1951).
Knowledge Representation’s appropriation of the ontology

Up until this point in our history, we have generally painted a picture of the ontology as an abstract model used to explain the nature of reality. However, with the introduction of the analytical approach and broad range of applications for this method proposed by the logical positivists, the principle that tangible problems could be resolved using logic-based models had been clearly demonstrated. Therefore, when the first computers arrived, the fact that many of their challenges came to be addressed using analytical approaches should not come as too great a surprise. Chief among these challenges was the need to allow computers to efficiently interpret information, and so important was this particular aspect of the computing question that it gave birth to its own sub-field of computer science, known as Knowledge Representation. KR came to view the ontology in far more concrete terms than had been the case up until that point. ‘The ontology moved from an uncountable noun in its philosophical sense to a countable noun in the computer science sense’ (Guarino et al. 2009 p. 1). The first definition of the ontology in the KR context to receive a sizeable support was proposed by Gruber in 1993. He defined the ontology as ‘an explicit specification of a conceptualization’ (T. Gruber 1993 p. 199). Borst expanded upon Gruber’s definition in 1997 by proposing that the ontology was ‘a formal specification of a shared conceptualization’ (Borst 1997 p. 12), and in doing so he introduced the idea that an ontology should be something that is shared among a group of agents, which, as we will see, was an important augmentation of the model.

It is on the in-depth and explicit study of the ontology’s definition prepared by Guarino et al. (2009) that we will now focus our attention. As mentioned above, Guarino’s ontology is very much a functional device to be used to achieve the objectives of KR; it is ‘a special kind of information object or computational artifact’ (Guarino et al. 2009 p. 2). And yet, it is also clear from Guarino et al.’s treatise that their understanding of the KR ontology is also very much grounded in a philosophical understanding of the ontology; the capitalised ‘Ontology’, as they would refer to it. The ontology in this sense forces the computer scientist to make explicit exactly what it is that they mean by a knowledge domain. Which concepts inhabit it? What are their attributes and what are their relationships to one another? And how might these concepts and their relationships be represented within a language that a computer can understand. We will now consider
how a KR ontology might be built to describe a set of activities at an archaeological site by applying the methods endorsed by Guarino and his colleagues.

**On the importance of Set Theory**

But before attempting this, we must briefly introduce the topic of set theory, first because it is the formal language used in the Guarino paper to describe the KR ontology but also, and more importantly, because it has become the principal method by which most KR research is currently undertaken. Set theory was first proposed as a mathematical logic system, which would be able to accommodate the concept of numerical infinity, which is a conundrum that has challenged philosophers since the time of the Presocratics (Dowden 2013). The model was proposed by the German mathematician, Georg Cantor in his 1874 treatise, Über eine Eigenschaft des Inbegriffes aller reellen algebraischen Zahlen (On a Property of the Collection of All Real Algebraic Numbers) (1874).

In this work, Cantor uses set theory as an alternative to the syllogisms, which had dominated logical reasoning since the time of Aristotle. The formal languages used by Frege, Russell and the analytical philosophers that followed them all owe an enormous intellectual debt to Cantor’s system. By the mid-20th century, when the first computer scientists were searching for a logic system with which to encode reasoning in machines, set theory proved an obvious choice, and, ultimately, all modern computer programming logic is fundamentally based on the set theory paradigm.

**Implementing the Guarino KR ontology model for archaeology**

Guarino et al. frame their definition of the KR ontology around three questions (Guarino et al. 2009 p. 3). What is a conceptualisation? What is a proper formal, explicit specification? And lastly, why is ‘shared’ of importance? For the first of these questions, let us imagine an archaeological site. At this site, there are three material entities (‘Frank Lynam’, ‘Joe Bloggs’, ‘Priniatikos Pyrgos’, or two people and an archaeological site) and there are two abstract entities (‘Person’ and ‘Site’). We can and should relabel these entities as concepts in order to conform to the terminology of conceptualisation. These are basically all of the concepts that I have formulated in my mind to represent the reality of this particular archaeological project. These concepts are related to each other in particular ways. For one thing, ‘Frank Lynam’ and ‘Joe Bloggs’ are related to the ‘Person’ concept, and ‘Priniatikos Pyrgos’ is related to the ‘Site’ concept. This conceptualisation domain can be represented schematically, as is shown in Figure 19.
Figure 19: a conceptualisation of the archaeological site knowledge domain

The important point to note about this conceptualisation is that it exists purely as an internalised model. It needs to be specified in some way in order for it to be realised, which will in turn allow it to be consumed by an agent other than its creator. This is where we need to start thinking in terms of the ‘specification’ requirement included in Gruber and Borst’s definitions. This language of the specification can be structured in various ways. To begin with, it could be in the form of natural language. If we were to follow this approach, our archaeological site example might be defined as is represented in Table 7.

<table>
<thead>
<tr>
<th>Concept</th>
<th>Relationship</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>'Frank Lynam'</td>
<td>'type' relation</td>
<td>'Frank Lynam' is of type 'Person'.</td>
</tr>
<tr>
<td>'Joe Bloggs'</td>
<td>works-at relation</td>
<td>'Frank Lynam' works-at 'Priniatikos Pyrgos'.</td>
</tr>
<tr>
<td>'Priniatikos</td>
<td>works-with relation</td>
<td>'Frank Lynam' works-with 'Joe Bloggs', and this is a symmetric relation</td>
</tr>
<tr>
<td>'Site'</td>
<td></td>
<td>because the order of the two concepts does not change the meaning of the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>relation.</td>
</tr>
</tbody>
</table>

Table 7: specifying an archaeological conceptualisation using natural language

This specification for the conceptualisation would work well if it were to be consumed by a human agent because most human brains are familiar with reading natural language. This ability would allow for the meaning contained within the specification to be
internalised and the conceptualisation understood. However, machines, for the most part, would find it difficult to extract the semantics of this piece of natural language. However, given the computer’s close developmental association with formal language, were the specification to be defined in these terms instead, the computer would be far more likely to be able to make sense of it. The above natural language specification could be translated into the set theory form shown in Table 8.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>R = {Person, Site, works-at, works-for}</td>
</tr>
<tr>
<td>3</td>
<td>Person(w) = {'Frank Lynam’, ‘Joe Bloggs’}</td>
</tr>
<tr>
<td>4</td>
<td>Site(w) = {'Priniatikos Pyrgos'}</td>
</tr>
<tr>
<td>5</td>
<td>works-at = {('Frank Lynam’, ‘Priniatikos Pyrgos’), ('Joe Bloggs’, 'Priniatikos Pyrgos')}</td>
</tr>
<tr>
<td>6</td>
<td>works-with = {('Frank Lynam’, ‘Joe Bloggs’)}</td>
</tr>
</tbody>
</table>

**Table 8: specifying an archaeological conceptualisation using natural language**

The meaning of this is intended to be the same (or at least very close) to that of the preceding natural language specification. In the language of set theory, we call these various declarations extensional because they list all of the elements and relations that occupy the set or the relation. In practice, however, it is very difficult or perhaps even impossible in certain circumstances to define the relations of a conceptualisation using the extension method. The alternative is to provide a logical definition, known as an intensional relation, that when computed will equate to the elements of the conceptualisation. Because of their rule-like form, intensional set theory statements are often referred to as axioms.

\[ D = \{x \mid x \text{ is an integer, } 10 > x > 0\} \]

An example of a very simple intensional set definition is shown above. It equates to and is the same as the set \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}. In some cases, however, it is not possible to produce intensional formulae, which can be used to fill a set or a relation. If we consider our example, it is difficult to imagine how we might populate our relations with members of the universal set without doing so manually. After all, these concepts are not numbers and do not follow that well-defined logical system. We cannot say, for instance that ‘Frank Lynam’ is greater than or less than any one of the other concepts.

---

65 If, for example, you wished to declare a set of all natural numbers, it would be impossible to do so using extensions.
Because of this restriction, it is the case that the vast majority of conceptualisations that deal with archaeological content need to be populated manually. Of course, once this initial work of population is completed, it is then possible to take advantage of the benefits that come with the use of a formal language approach to conceptualisation and we will show how these types of analyses can be put into action later in this chapter.

**Figure 20:** the journey of real phenomena from reality to ontologies (after Guarino et al. (2009 fig. 2))

But for now let us return to Guarino et al.’s model of the KR ontology. The schematic shown in Figure 20 describes the various stages of the model. As we have seen, the first stage is concerned with the perception of real ‘phenomena’ and their internal conceptualisation. We also looked at how this conceptualisation comes to be specified using a formal language. In this context, Guarino and colleagues introduce the philosophical idea of an ontological commitment (Church 1958), which simply means that in order for an agent to engage with a conceptualisation, they must adhere to its conventions, the first of which is the form of its representation. In other words, in order
for a conceptualisation to be discussed by two agents, they must both be proficient in the use of the same language.

Borst’s reference to sharing is also important in this context. The specification of a conceptualisation must be by definition a social construct. If it is not, it has no meaning. Figure 20 also highlights how a conceptualisation can be specified in many different ways. After all, it is an objective interpretation of a subjective set of beliefs. Guarino’s diagram shows how one specification can be deemed ‘better’ than another, although in practice this rating could only be based on subjective value-judgements.

**KR ontologies in the Semantic Web**

A short introduction to Semantic Web ontologies

http://linkedarc.net/ontologies

In order to transform the archaeology site conceptualisation described in the last section into a machine-readable ontology we need to select a suitable formal language. We have shown in Chapter 2 how RDF is used to serialise triple information that is to be published to the Semantic Web. As it happens, ontologies, which structure these RDF datasets can also be formally represented using RDF in partnership with a sort of meta-ontology, which is a set of vocabulary terms that allows the ontology designer to define its reasoning. For example, an ontology implementation needs to be able to declare concepts. It also needs to be able to declare the relations that link these concepts. And it might also want to be able to define an internal hierarchy for these concepts and relations. Most of these needs should be familiar by now, as they also define the Guarino conceptualisation. The only major difference between defining the rules of an ontology using the language of set theory and doing so using RDF and a meta-ontology is one of terminology. These meta-ontologies for the most part use the terms ‘class’, ‘property’ and ‘taxonomy’ to refer to the ‘concept’, ‘relation’ and ‘hierarchy’ used in the Guarino definition. Let us now consider how we might specify the earlier archaeological site example using one of the most common RDF meta-ontologies: Web Ontology Language or OWL.
Ontology authoring using OWL

In 2004, the W3C released a recommendation for an RDF vocabulary set called RDF Schema or RDFS, which allowed users to create basic ontologies using an RDF data representation base (Brickley et al. 2004). In the same year, the organisation published another recommendation for a comparable but far more expressive vocabulary,\(^6\) which they called Web Ontology Language or OWL (McGuinness & van Harmelen 2004). Both have since been updated, with a version 1.1 of RDFS being released in 2014 (Brickley et al. 2014) and OWL following suit by providing a version 2.0 update in 2014 (Hitzler et al. 2012). OWL comes in a number of different configurations, each of which is designed to cater for a different level of ontological complexity. All of these OWL variants were designed to subsume the functionality of RDFS and, as such, most OWL implementations include vocabulary taken from RDFS.

For our archaeological site example, the first task that we need to accomplish is to select the classes that we will need to express the meaning contained in our conceptualisation. In the Guarino model, we defined five concepts, ‘Frank Lynam’, ‘Joe Bloggs’, ‘Priniatikos Pyrgos’, ‘Person’ and ‘Site’ to describe the particular knowledge domain. A Guarino concept does not, however, map in a one-to-one fashion onto an OWL class. An OWL class is an abstract entity, which models a particular type of concept and in that sense is analogous to a Platonic form. OWL instances, on the other hand, are realisations of that abstract class (Antoniou & Van Harmelen 2004 p. 81) and, as such, can be compared to a Platonic particular. Following this distinction, only the Person and Site concepts qualify for the status of class in our OWL ontology. The remaining concepts are to be implemented as instances of these two classes, and we will deal with the creation of these instances later.

```r
@prefix : <http://example.org/ontology/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@base <http://example.org/ontology> .
:Person a owl:Class ;
    rdfs:label "Person class"^^xsd:string ;
```

\(^6\) See Antoniou and Van Harmelen (2004 pp. 111–112) for a list of the various limitations that RDFS imposes on the ontology designer.
Following the convention of this thesis, we will implement the OWL ontology for our archaeology site example using RDF Turtle, as is shown in Table 9. As you can see, there is nothing unusual about this Turtle serialisation. It includes a header, which outlines the various prefixes used throughout the document and a body section, which contains groups of RDF triple declarations, which share the same RDF subject. The first pair of these triple groupings defines the characteristics of our two ontology classes, Person and Site.

The Person class declaration (lines 8-13) begins with a statement declaring that a subject with the URI <http://example.org/ontology/Person> is of type owl:Class. This means that we are defining <http://example.org/ontology/Person> to be a class in our ontology. Line 9 declares a label for the class and lines 10-11 a comment. It is important to describe the classes that are declared in an OWL ontology. One of the most important functions of an ontology is to render transparent the structure of the dataset that it is
applied to. It is not a requirement to include this metadata in the class definition but it helps enormously when the time comes to share the ontology. Line 12 uses the owl:disjointWith predicate to declare that Person class instances cannot also be Site class instances. owl:disjointWith illustrates how ontologies are essentially sets of rules or axioms by which ontologically committed agents must abide. This tells the creator of the dataset that applies this ontology that they cannot declare an instance that is a Person and a Site. This would be an illogical or meaningless statement. The same is true for the data consumer. The ontology informs the consumer that they can take it as read that instances cannot be a Person and a Site at the same time. The final triple declared in this group (line 13) fulfils another administrative duty. It states that this class definition resides within the <http://example.org/ontology/ArchaeologySiteOntology> ontology scheme. Lines 15-20 define a similar set of rules, which govern the logic of the Site class.

A conceptualisation does not just define the concepts in a knowledge domain. It also describes the relations between those concepts. In the archaeology site conceptualisation, two relations were described, one to define a co-working relationship between two Person instances and another to declare that a Person instance works in the location of a Site instance. These relations are known as properties in OWL and they come in two varieties: object properties and data properties (Antoniou & Van Harmelen 2004 p. 118). Object properties relate class instances to one another and data properties relate a class instance to literal values.67

![Figure 21: defining an ontology relation using OWL](image)

Lines 22-29 in our Turtle serialisation define the first of our ontology’s properties. It is an object property, as it links one class instance to another. Therefore, it is of type owl:ObjectProperty. A number of the triples in this group have been described in the context of the class definitions, and they have the same meaning here. The rdfs:domain, rdfs:range and rdfs:subPropertyOf predicates are specific, however, to property definitions. rdfs:domain defines the RDF triple subject that can use this property as a

---

67 For an account of RDF literal values, refer to Chapter 2.
predicate and rdfs:range defines the RDF object that the subject can link to. rdfs:domain and rdfs:range essentially restrict the types of subjects and objects that can create an RDF triple around this property (Figure 21). rdfs:subPropertyOf is used to place the property within a hierarchy of properties. In the case of the <http://example.org/ontology/works-at> property, line 28 declares that the property is the child property of owl:topObjectProperty, which is the most abstract property available within the OWL vocabulary. Declaring property and class hierarchies allows an ontology-modelled RDF dataset to be subjected to advanced forms of logical reasoning.

There are many other OWL vocabulary predicates, which can be used to introduce more advanced forms of reasoning into an ontology. For instance, it is possible to associate the meaning of two classes using the owl:equivalentClass predicate or to implement a hierarchy of classes using the rdfs:subClassOf predicate. owl:equivalentProperty associates two properties together and a property can be associated with the inverse meaning of a second property using the owl:inverseOf predicate. More specific types of properties can also be defined using OWL. For instance, it is possible to define a property as transitive, symmetric or functional (can have only one value) using the owl:TransitiveProperty, owl:SymmetricProperty and owl:FunctionalProperty classes.68

**Applying a Semantic Web ontology**

Once an ontology has been created, it can then be used to model real data. The conceptualisation proposed for the archaeological site example included a number of concepts, which were not turned into OWL classes. These are what are known in the KR world as instances or objects and they are related to the abstract classes that they realise. Let us now consider how we might implement our archaeological site ontology by creating these instance concepts.

The serialisation shown in Table 10 completes the internalised conceptualisation for the archaeological site example by applying the ontology that we built to model an RDF dataset. We declare the resource <http://example.org/data/person1> (lines 7-10) to be of type archsiteont:Person, which as you will remember, is the OWL class that was declared and described in our ontology. This creation of instances is often called

---

68 We could for instance have declared the <http://example.org/ontology/works-with> to be of type owl:SymmetricProperty because its subject and object can be interchanged, and the meaning remains the same.
‘instantiation’. The triples for this resource utilise both the archsiteont:works-at and archsiteont:works-with property also defined in the ontology. Notice how the domains and ranges of these triples obey the rules as set out in the ontology. For instance, the archsiteont:works-at property has a Person instance as its subject and a Site instance as its object.

```xml
@prefix : <http://example.org/data/> .
@prefix archsiteont: <http://example.org/ontology/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@base <http://example.org/data> .

:person1 a archsiteont:Person ;
    rdfs:label "Frank Lynam"^^xsd:string ;
    archsiteont:works-at :site1 ;
    archsiteont:works-with :person2 .

:person2 a archsiteont:Person ;
    rdfs:label "Joe Bloggs"^^xsd:string ;
    archsiteont:works-at :site1 ;
    archsiteont:works-with :person1 .

:sitel a archsiteont:Site ;
    rdfs:label "Priniatikos Pyrgos"^^xsd:string .
```

Table 10: instantiating the archaeological site ontology

Were this dataset to be published to a triplestore with support for SPARQL, it could be interrogated using the query shown in Table 11. This would produce the result shown in Figure 22.

```sql
PREFIX archsiteont: <http://example.org/ontology/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT * WHERE {
    ?person1 a archsiteont:Person .
    ?person1 archsiteont:works-at ?site .
    ?person1 archsiteont:works-with ?person2 .
    ?person1 rdfs:label ?person1Name .
    ?person2 rdfs:label ?person2Name .
    ?site rdfs:label ?siteName .
}
```

Table 11: a SPARQL query addressed at the archaeological site RDF dataset
Chapter 3 - Data transparency and re-use

**On advanced machine reasoning**

To date we have presented a picture of the KR ontology as a formalisation of a conceptualisation, which allows a machine to interpret its meaning and enables the conceptualisation to be shared among a wide agent base. It has also been shown that the authoring of ontologies does not assume a narrow outcome – multiple ontologies can be created to model the same conceptualisation, and because a conceptualisation is by definition a subjective perception of a natural phenomenon or set of phenomena, it also stands to reason that more than one conceptualisation can be derived from this source.

In summary, the creation of ontologies is complex and as a general rule, the more effort that the data designer puts into this process, the more analytical and interpretive gains can be enjoyed when it is put into use.

We have talked about the value of the ontology as a sort of semantic glue in the building of RDF datasets. We have also seen its use in educating consumers of those datasets. While these are clearly valuable outcomes and perhaps the primary usefulness of the ontology in the context of its use within the Semantic Web environment, these are not the only function that it can perform. Ontologies can also be employed to support interrogations of datasets that share the reasoning load between the human and the machine agents. For example, let us imagine that it is your intention to search across two separate datasets, which both host data about two local assemblages of amphorae. Each dataset has created its own ontology to model this data but these custom ontologies have been based on a single public ontological source, which deals in more general terms with ceramic concepts. This is a fairly typical scenario. Often custom ontologies will use a public and well-known ontology on which to base their logic. Usually, they will sub-class a number of the public ontology’s classes and properties. The objective here is to model the specificity that differentiates their data from other data and which gives it its unique value, while at the same time, they are able to profit from the logical robustness and visibility of the tried and tested model.
For the agent (human or machine) approaching these two amphorae datasets for the first time, they will first want to ask questions at a more generic level (Figure 23). They know that both datasets’ ontologies derive from a public ceramics ontology, whose structure is familiar to them. Because of this knowledge they are able to ask both datasets for a list of all ceramic vessels in their respective collections. Even though the datasets have only declared instances of their own custom classes, which represent their own conceptualisation of the amphora entity, because these classes extend the public ceramic
vessel class, the inferencing engine of the RDF triplestore, which houses the data, is able to reason that the response to the agent’s query must also include instances of the custom amphora classes. Essentially, this is like asking a zoo for a list of all of its animals. Because giraffes, lions and bears are all defined as sub-classes of the animal concept, the zoo will include these more specific classes of animal in its response. This reasoning method, which we call inferencing, seems to make common sense to a human because it is one of the basic ways that we reason through conceptual problems (J. S. B. T. Evans et al. 1993 pp. 1–2). In our example, nowhere in the two datasets is it explicitly stated that these amphora instances are also instances of the ceramic vessel class. However, because this semantic information is included in their ontologies, the reasoning engine of the triplestore is able to make this semantic connection.

Another advantage of using machine reasoning becomes apparent when you consider the problem of merging the ontologies of multiple datasets during data aggregation. Perhaps, several previously distinct datasets need to be amalgamated into the one store. This would require their mapping from several ontological bases to a single shared ontology (de Bruijn et al. 2006 pp. 8–10). Often these mappings will involve a generalisation of the semantics for all of the datasets involved, reducing their meaning down to a common semantic denominator. This generalisation process results ultimately in a loss of the semantic richness of the combined dataset. There is, however, an alternative. OWL provides the property, owl:unionOf, which allows new classes to be created in the amalgamated ontology that combine the meanings of the source classes. If, for example, we wished to aggregate our two amphorae datasets into a single ontology and dataset, we could define a new class called GeneralAmphora that would be a union of the two specific class definitions, as is shown Table 12.

<table>
<thead>
<tr>
<th></th>
<th>:GeneralAmphora owl:unionOf [:Dataset1Amphora, :Dataset2Amphora] .</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Table 12</strong>: combining two ontologies using OWL</td>
</tr>
</tbody>
</table>

Lastly, the SPARQL language itself is an example of advanced logical reasoning in action. For instance, it is possible to process the results generated by a SPARQL query before they are sent back to the client. Using the SPARQL aggregate function, COUNT, for example, allows you to have the SPARQL engine count up the amount of responses in any one variable and to have that value returned to the querying agent. SUM adds up the
integer values of a set of results, and MIN and MAX look for the lowest or highest value in a set of integers.

**Semantic Web controlled vocabularies and SKOS**

A dataset type, which is strictly speaking not an ontology but is a related form in the sense that it provides ontological services to an RDF dataset, is the controlled vocabulary, and its instances are encountered regularly on the Archaeological Semantic Web. Also referred to as a thesaurus or taxonomy, the controlled vocabulary is a specification of a conceptualisation that is structured hierarchically and is used primarily to provide sets of terms, which can be employed as object values in RDF graph data. In this section, we introduce and describe the form and function of the controlled vocabulary, looking specifically at how the SKOS model can be used to create Archaeological Semantic Web controlled vocabularies.

![Figure 24: Ramon Llull's Tree of Science (Lima 2011 fig. 11)](image)

The tree form has a long and established history as a conceptualisation and visualisation motif in both the sciences and the humanities. The tree of life, for instance, is a metaphor that is often used to explain the structure of aspects of the natural world. Hagedemer (2008 p. 8) says that the tree is ‘an image of the whole universe, or at least of our planet, that embodies the notion that all life is interrelated and sacred.’ A representational model that highlights the interrelatedness of things is not only useful to
explain the natural sciences. Many aspects of knowledge can be viewed in similar terms. The bible divides knowledge up into what it deems to be good and evil branches of the one tree (Lima 2011 p. 25). The tree form was also used extensively throughout the Middle Ages, as is illustrated in Ramon Llull’s magnificent 13th century Arbor scientiae (Tree of Science) image, which conceptualises the world of science as a tree of eighteen roots representing the philosopher’s principles and sixteen branches representing the different domains of science (Figure 24).

Certain types of information lend themselves to be conceptualised in the arboreal form more than others. Concepts that form part of a tree hierarchy can only be related in one of two ways. They can be the child of a concept and/or the parent of a set of concepts. Compared to the potential complexity of a full-blown ontology, in which two concepts can be related in a host of different ways, each carrying a different meaning, the tree model is markedly simpler. In the Semantic Web, controlled vocabularies are almost always implemented using the Simple Knowledge Organization System or SKOS, which was formally recommended by the W3C in 2009 (Isaac & Summers 2009; Miles & Bechhofer 2009). SKOS, as the name intimates, is a straightforward model. The ontology is built around a single class, skos:Concept. Builders of controlled vocabularies are able to create instances of this class, give them a name and link them to other skos:Concept instances in their or other controlled vocabularies in order to create a hierarchy of concepts. The objective is to use these skos:Concept instances to populate the objects of RDF triples in other datasets.

As with all types of Semantic Web ontologies, it is far preferable to reuse existing concepts if they are found to already exist. In order to encourage the re-use of SKOS controlled vocabularies, an emphasis is placed on the descriptive quality of the concepts that it models (Mader et al. 2012 pp. 226–227). In this context, so-called ‘scope notes’ play an important role as they outline in descriptive terms the intended meaning of any one concept. Of course, this does not remove the potential for semantic ambiguity but it certainly helps to clarify the intent behind a concept’s creation.

---

69 Juan-Antonio et al. (2009) document some of these alternatives.
Table 13 shows the Getty Art and Architecture Thesaurus (AAT) SKOS declaration for its silver metal concept. The resource is first declared as an instance of the skos:Concept type (line 1), and it is then assigned a label in a number of different languages (lines 2-4). Three broader concepts are linked to the concept on lines 5-7. This implies that the AAT’s silver concept is part of three separate knowledge trees (a default knowledge tree, an artefact quality knowledge tree and an artefact material composition knowledge tree), with each of the listed resources being a parent of the silver concept in the respective tree. The skos:inSchema property associates the resource with the AAT dataset (line 8), while the skos:narrow property identifies a single child concept, which represents the silver alloy concept. The skos:related property links the resource to other miscellaneous related concepts, such as silverwork, silverware and silversmithing (lines 10-15). Lastly, the skos:scopeNote triple defines its designers’ intended meaning for the concept (lines 16-25).

---

This representation was retrieved from the Getty server on 23 July 2015. Note that a number of the non-SKOS-related triples have been removed from this reproduction. The Getty RDF datasets are discussed in Chapter 4.
Chapter 3 - Data transparency and re-use

The resources that are defined within SKOS datasets are not classes; they are instances of the skos:Concept class. As such, a dataset that wishes to use one of these concepts does not link to the resource using the rdf:type predicate, as would be done if the SKOS instance were a class. Instead, a SKOS concept must be linked to a resource using another property, which is determined by the ontology that the dataset conforms to. For instance, the Dublin Core’s dc:subject property is often used to associate a resource with a SKOS concept, and as we will see in the next section, the CIDOC CRM provides its own analogous property, ecrm:P2_has_type.

**Archaeological Semantic Web ontologies**

Whether it be in the refactoring of data within existing datasets or the use of Open-Source Software infrastructures to host Archaeological Semantic Web content, the benefits from a collective recycling of approaches is championed in all corners of the Archaeological Semantic Web community. In no aspect of the Archaeological Semantic Web’s operation is this principle more strongly encouraged than for ontologies. In this section, we consider the CIDOC CRM model as an example of how this domain-wide ontology paradigm can work in practice.

**The CiDOC CRM**

[linkedarc.net resource](http://linkedarc.net/ontologies/ecrm)

Practical notes on implementing the CIDOC CRM using RDF

In Chapter 2 we reviewed the importance of the CIDOC CRM in the history of archaeological and cultural heritage information systems. The CRM abbreviation stands for Conceptual Reference Model. At one level, this can be viewed as largely analogous to the ontology as understood in the KR sense. However, it also refers to the fact that the CRM is intended to act as a guide for how cultural heritage digital datasets should be constructed and communicated across networks of knowledge exchange. Until 1994, the CRM was based on an entity-relationship model. In that year, the decision was made to overhaul the model’s basic structure by moving to an object-orientated design. The first instalment of the object-orientated version of the CRM was released in 1998 and since then it has gone through five major iterations, as well as being released as a full ISO

Despite the intervening 17 years, the model’s core structure and ontological aims have remained largely unchanged. An early document outlining the position of the CRM identifies its primary role ‘as being to define a semantic framework which will enable compatible [cultural information] systems to exchange and share information’ (Doerr & Crofts 1999b p. 2). This document also states that links ‘to other objects, and to an object’s historical, geographical, and cultural origins help to place it [the cultural heritage digital resource] in a context and give it meaning’ (Doerr & Crofts 1999b p. 3). Doerr and Crofts note that the combining of ‘information from different sources requires a high level of abstraction and a discipline neutral viewpoint, which has the flexibility for different viewpoints to be respected and expressed. This generic level of abstraction is precisely what the CRM aims to provide’ (Doerr & Crofts 1999b p. 4). This is a key statement, as it defines the primary intellectual approach taken by the CRM. It is both the model’s greatest strength and in the view of its numerous detractors, its greatest deficiency, as we will see.

**The basic design philosophy**

The idea of conceptual abstraction is fundamental to the CRM model. The CRM does not dictate to cultural heritage digital providers on how to document their data but rather it assumes that all documentation is conceptually the same at a certain generic level. As such, the CRM’s role is to highlight just how an existing model sits within this universal conceptual framework (Le Boeuf et al. 2015 p. i). If we suspend our misgivings about the obvious ontological and epistemological problems that this assumption introduces, the challenge of the CRM is, therefore, to create a level of abstraction that can service the largest possible need. Or in other words, the CRM needs to guarantee the most expansive ontological commitment (Doerr 2003 p. 5).
Figure 25: relations between the core CRM classes (Doerr 2003 fig. 3)

To realise this abstraction, the CRM defines three main classes: Actor, Event and Object. Although many other classes are also defined by the ontology, the majority ultimately derive from these classes. The general pattern of class interactions allowed for by the CRM is described schematically in Figure 25. The CRM is positioned as a reaction against object-centric models, such as the Dublin Core, which promote the modelling of reality around an object referent and which tend to produce single perspective datasets. As an alternative epistemological structure, the CRM instead adopts an event-based model (Oldman & Labs 2014 p. 9), which allows for the modelling of multiple representations of the one real phenomenon.\(^\text{71}\)

An example, often used to explain the CRM process, is the modelling of artefacts associated with the 1945 Yalta Conference, at which an agreement to end the hostilities of WWII was signed (Morray 1974). The artefacts modelled are the actual agreement signed at the conference, which is held in store by the US State Department and a digital representation of a famous photograph of the leaders at the event, which is stored on the Wikimedia servers. And then there are the contextual elements, some real and some conceptual, such as the location of Yalta, which is found in modern Crimea, the date of the conference and the individuals involved: Churchill, Stalin and Roosevelt. A graph describing this network of relationships is shown in Figure 26. At the centre of this network is an instance of the E7_Activity class, which is a sub-class of the E5_Event

\(^\text{71}\) For more on the implementation of multiple interpretations using the CRM, see Chapter 7.
class and which represents the conference event itself. This conference activity contains another sub-type of activity, which is modelled using the E65_Creation class and whose instance represents the creation of the Yalta agreement. The Yalta modelling also includes instances of objects in the form of the E38_Image of the leaders and the E31_Document, which represents the agreement’s physical document. It also includes an instance of E53_Place, which models the place of Yalta. The dates of both the conference holding and the signing of the agreement are also represented using the E52_Time-Span class. Lastly, the three leaders are represented as instances of the E39_Actor class.

Figure 26: a graph modelling the signing of the Yalta Agreement using the CRM (Sinclair 2015)

All of these various instances are linked to one another via CRM properties and the CRM specification outlines the formal rules that govern the domains, ranges and quantifications of these properties. For instance, the P4_has_time-span property can be employed to link a subject of type E2_TemporalEntity and an object of type E52_Time-Span. The specification also states that P4_has_time-span can be used in many-to-one relationships. This means that more than one subject can link to a single object using this property, which makes sense, as a time-span instance is unique, while its subject might be any number of instances. On the other hand, many of the CRM properties allow for more flexible relationships, which in turn can be used to model multiple truths into a dataset. For example, the P7_took_place_at property is a many-to-many property. This
means that it is possible for one interpretation to say that the Yalta conference E7_Activity took place in E53_Place Yalta and for another to disagree with the first interpretation by including a link to a different E53_Place instance.

Because the labels and semantics of the CRM classes are left as generic as possible, many cultural heritage data modellers have contended that the model is overly esoteric and complicated to understand and implement (Carver 2013; E. C. Kansa 2014). There is, for instance, no CRM class, which more specifically models the people that occur in the example. They must all be defined simply as instances of the E39_Actor class. In many cultural heritage information systems, people will be defined in more domain-specific terms, for instance as human bone specialists or as English language literature specialists. The CRM makes no accommodation for these types of specialist needs and in many cases this means that the CRM needs to be extended if it is to be implemented at all, and this is the subject that we turn to next.

**Archaeological extensions of the CRM**

The CRM is declared by its authors to be an ontology aimed at servicing the digital needs of the museum sector (Le Boeuf et al. 2015 p. i). However, because of the recent trend towards the amalgamation of different information sources, which are defined to be part of the same general knowledge domain, and which is manifested most clearly in the form of the Semantic Web, it has become increasingly apparent that the CRM classes and properties need to be extended so as to accommodate these differing needs. Thankfully, the object-orientated CRM was designed with the idea that it would be extended upon (Le Boeuf et al. 2015 p. xvii), and a number of extensions have already been developed to meet such various needs as canonical citation (Romanello & Pasin 2011), biographical modelling (Reinert & Thomas 2011) and geospatial referencing (Hiebel & Doerr 2013). And, as we have seen in Chapter 2, the CRM has also been extended twice to produce conceptual models for the structuring of archaeological data. We will now briefly introduce the design of the Single Context-based CRM-EH extension and then we will consider the more recent and more methodologically neutral CRMarchaeo extension.

---

72 Chapter 4 introduces a number of these cultural heritage data aggregation projects.
The CRM-EH model contains a total of 126 new classes, seven new object properties and five new data properties, all of which derive from CRM core definitions and are modelled using OWL (Binding 2014a). The breadth of the CRM-EH is illustrated in Figure 27, which shows how the various classes relate to one another. It would be inaccurate to say that the CRM-EH is an overly complex model (at least, it is no more complex than its ontological parent), but it is certainly expansive, and the primary reason for this is that it attempts to model such a large swathe of the Single Context (SC) archaeological process. The model is divided up into different zones, each representing a particular type of activity (surveying, project initiation, drawn records, the site, photo record, finds dating, finds documentation, finds processing and analysis, archaeological
science, context record, documentation of contexts, context excavation event, grouping and phasing) carried out as part of the SC process.

Within each of these separate groupings, there is a microcosm of the CRM event-based model in action, with each CRM-EH class being ultimately derived from the CRM Actor, Event or Object class. For example, a find is modelled as an EHE0009_ContextFind (derived from the CRM E70_Thing class) instance. This is given a unique identifier in the form of an EHE0043_ContextFindUID (derived from the CRM E70_Thing class) instance via an EHE2013_ContextFindIdentifierAssignmentEvent (derived from the CRM E5_Event class) instance. The classification of the find is carried out by an individual, who is modelled as an EHE0077_ProjectTeamMember (derived from the CRM E39_Actor class) instance.

When first approached, the CRM-EH’s modelling of the context concept can be somewhat confusing. The reason for this is that the CRM-EH divides the conceptual complexity of the context across a number of different classes. First, there is the EHE0007_Context class, which is a sub-class of the CRM E53_Place class. This models the context in its understanding as the space within which the material of the context is located. This material is modelled using the EHE0008_ContextStuff class. Then there is the idea of the context as a discreet event in time that caused the creation of this material context – for example, a pit is dug into the ground. This context is modelled using the EHE1001_ContextEvent class. The actual excavation of the context also needs to be modelled and again this gets its own class, EHE2001_ContextExcavationEvent. Finally, there is the EHE0012_ContextEventRecord class, which derives from the CRM’s E73_Information_Object class, and is used to represent the information artefacts (record sheets, plans, et cetera) used to record the context during excavation and in post-excavation.

Clearly, the CRM-EH context is no simple ontological matter. Perhaps, this approach to the context is more reflective of the actual complexity of what is such a fundamental concept within archaeology. On the other hand, ontological honesty can come at the price of practical confusion, and it is likely that the multifarious understandings of the context by the CRM-EH have posed a problem for many potential adopters.
**CRMarchaeo**

The CRMarchaeo extension is still very much in its gestation period. The current version, 1.2.1 and the release, on which we will base this following account of the model, was released in October 2014 (Cripps et al. 2014). The extension’s objectives are comparable to those of the CRM-EH, but they adopt a somewhat broader perspective. Whereas, the CRM-EH was created to model only data derived from SC excavation, the CRMarchaeo extension is designed to model the processes and output of any archaeological activity carried out in Europe (Masur et al. 2014). The model’s reference documentation adds that the CRMarchaeo aims to ‘maximise interpretation capability after excavation’, to allow for the ‘possibility of knowledge revision after excavation’, to compare ‘previous excavations on the same site’ and to accommodate ‘all kinds of comprehensive statistical studies’ (Masur et al. 2014 sec. 1.1.1).

As well as having similar objectives, the ontological structures of the two CRM extensions are also very alike, although the language used to describe the CRMarchaeo is more generic, speaking of ‘stratigraphic units’ and ‘excavation processes’ where the CRM-EH referenced ‘contexts’ and ‘context excavation events’. Essentially, both models address the archaeological process by dividing it into those concepts, which are related to excavation on the one hand and those, which we can define as interpretive on the other. Of course, it is easier to hypothesise about such a separation than it is to actually achieve in practice. In recognition of this fact, the CRMarchaeo admits that the best that can be hoped for may be in the creation of more detailed archaeological paradata, which can then be fed into the archaeological interpretive process (Masur et al. 2014 p. 9).

In this respect, CRMarchaeo also declares more than one class to model what they call the stratigraphic unit (there are four out of a total of eight classes, which deal with stratigraphic concepts), as well as including a single class, A1_ExcavationProcessUnit, to model the excavation process. For the most part, the logic of CRMarchaeo comes across as being far more robust and approachable than that of the CRM-EH.73 The schematic shown in Figure 28 illustrates how the CRMarchaeo’s designers intend for its classes to be implemented. There is a distinction made between a stratigraphic interface (A3_StratigraphicInterface) and the material bounded by these interfaces

---

73 A number of the individuals who were involved in the building of the CRM-EH are also now involved with the ARIADNE project. It is likely, therefore, that the experience of the CRM-EH has greatly informed the development of the CRMarchaeo.
Chapter 3 - Data transparency and re-use

(A2_StratigraphicVolumeUnit), and both of these classes derive from the parent A8_StratigraphicUnit class. The act of excavation is represented by the A1_ExcavationProcess class and Figure 28 shows how it is possible to model any form of excavational methodology using this class; in one case, the A1_ExcavationProcess models an arbitrary ‘spit’ excavation of the area, and in another, it models the employment of a stratigraphic excavation method.

Figure 28: an example showing the mapping of the CRMarchaeo classes onto archaeological phenomena (Cripps et al. 2014 fig. 4)

Instances of A8_StratigraphicUnit, A3_StratigraphicInterface or A2_StratigraphicVolumeUnit can all be related chronologically to other instances of the same type, thereby allowing Harris Matrix-type analyses to be carried out on a modelled dataset. Figure 28 also includes an example of an A7_Embedding class instance in the form of a coin find. Again, it is the case that where the CRM-EH had a multitude of classes to model a range of concepts, the CRMarchaeo has decided to adopt a less specific and less fragmented approach.

Structurally speaking, the CRMarchaeo actually extends an intermediary model, which for its part extends the CRM. The CRMsci is a CRM extension, which was created to model data that is created by the application of the empirical scientific method (Doerr et al. 2014). For example, the CRMsci includes classes that model concepts such as an empirical observation, a material form or part of that form, material quantities and the events that bring a material into being and that alter it. The authors of the CRMarchaeo must have decided that this pre-existing concept set could be used as a conceptual basis.
on which to model the archaeological process. Certainly, archaeological positivists would agree with this position but it would undoubtedly present more ontological and epistemological challenges for a post-processualist. This is an example of how the modelling of data is inherently a philosophical and indeed even a political act, and it is a subject that we will return to in Chapter 7.

**Conclusions**

This chapter has presented the Semantic Web ontology as the latest iteration in a long tradition of philosophical questioning. Superficially, one might think that a lot has changed since Parmenides wrote his sacred poem about his transcendental journey towards an understanding of the nature of being. After all, over two millennia later, the Vienna Circle’s approach to philosophy would appear to be as far removed from the 5th Century mystic’s method and views, as is possible to contemplate. And yet, on closer inspection, it could be argued that the Vienna Circle’s championing of empiricism is not that unlike the Aristotelian approach to philosophy. And similarly, one might suggest that Hegel and Kant’s German Idealism owes its own intellectual debt to Parmenides, Pythagoras and many of the other Presocratics. Philosophical positions, it seems, can oscillate through time and in so doing they serve to alter our shared ontological models.

Today, ontologies determine such a vast proportion of the workings of the information systems that we depend upon as scholars of the digital age and yet despite this significance, they are often viewed solely as the objective mechanisms through which the real subjective work of the scholar is achieved. As we have seen in this chapter, nothing could be further from the truth. The ontology, in the context of its employment within the field of KR, is a meaningful, deterministic and if used ignorantly, detrimental variable in the 21st century knowledge creation process. While the tendency is to ignore this fact, Guarino’s work on the subject is an example of what can be achieved when a more reflexive perspective is adopted.

We also looked in this chapter at the structures and reasoning behind the ontologies that currently enjoy the greatest following within the archaeological and broader cultural heritage fields. The CIDOC CRM has now assumed a hegemonic status within this space, but despite its popularity, the battle to win over the hearts and minds of the broader archaeological community still remains largely to be won. The recent release of the CRMarchaeo extension by the ARIADNE project can be seen as an attempt to
redress this balance. First impressions of the CRMarchaeo are promising. The model is markedly simpler in structure than previous efforts to model archaeological data, and presumably, this is a very deliberate design feature of the ontology.
Chapter 4 – linkedarc.net
On building an Archaeological Semantic Web RDF triplestore

‘While we’re acknowledging writing theory as making stuff, can we also acknowledge making stuff as doing theory?’ (Kirschenbaum 2013)

Introduction
Methodological knowledge, or the principle that there is epistemic value to be found in practice, is a central idea in the digital humanities (McCarty 2005 p. 120). Archaeologists, because of the focus that they place on material things, are comfortable with this concept. Branches of the discipline, such as experimental archaeology (Outram 2008) and material cultural studies (DeMarrais et al. 2005), and popular archaeological schools of thought, such as phenomenology (Tilley 1994), all see the value in the act of doing. While the tools of the trade of the digital humanist tend to be different from those of the archaeologist, the principle remains the same. In place of the trowel and the drafting board, the digital humanist employs the use of the keyboard, the digital display screen and the computer. This chapter presents the central practical component of this thesis, that of the building of linkedarc.net, an Archaeological Information System. I have designed linkedarc.net from the outset to be fully compliant with Linked Open Data practice, insofar as has been practically possible and achievable within the scope of this project.

No programme of work exists in a knowledge vacuum, however, and we begin this chapter with a review of Archaeological Information Systems and Cultural Heritage Information Systems, which are noteworthy for their adherence to Linked Open Data practices. Having situated linkedarc.net within its technological and intellectual space, we move on to the primary topic of the chapter. This involves an in-depth account and explanation of the linkedarc.net server and web app technical design.

Linked Open Data and archaeology: the trend setters
The archaeological method is fundamentally built on an ability to assimilate, categorise, analyse and interpret large amounts of disparate data with a view to ultimately finding meaning in them (Hodder & Hutson 2003 p. 209). In the following section we analyse a
number of projects that are applying this philosophy of practice using techniques that are compatible with the Archaeological Semantic Web ideal.

The Numishare framework, Nomisma and the Online Coins of the Roman Empire

The Numishare framework was developed by Ethan Gruber and Andrew Meadows of the American Numismatic Society as a software platform on which to allow the creation, management and dissemination of numismatic data online (E. Gruber 2015). The project blossomed out of Gruber’s efforts to make the University of Virginia’s Art Museum’s Numismatic Collection available as an online resource (University of Virginia Library 2015). The figurehead of the Numishare community of applications is Nomisma. Nomisma is primarily a Linked Open Data vocabulary of terms, which can be used to populate the data fields of digital coin collections (E. Gruber 2012; Meadows & Gruber 2014). In this role, Nomisma acts as a collection of ‘minted’ URIs that can be used by coin databases, thereby making them compliant with Berners-Lee’s fourth law of Linked Data (Berners-Lee 2006). While Nomisma primarily holds URIs for Roman Republican and Imperial coin types, mints, persons related to the creation of coin material and emperors and deities represented on numismatic iconography, it also, but to a far lesser extent, hosts URIs for corresponding entities for ancient Greek coin types, which are decidedly less canonical.

Nomisma also performs a secondary valuable function and that is to aggregate numismatic content, which has been created by its partner projects, which are not insignificant in number. Principal among these has been the American Numismatic Society and again building on the Numishare framework and the Nomisma URIs, Gruber and the Institute for the Study of the Ancient World at New York University have produced the Online Coins of the Roman Empire (OCRE) collection (American Numismatic Society & Institute for the Study of the Ancient World at New York University 2015). OCRE is an impressive achievement, albeit one with a few caveats. It provides the user with a faceted search function, which makes the discovery and analysis of data on the site intuitive and powerful. It is also possible to generate graphs that summarise certain aspects (the percentage of certain coin types for example) of the

---

74 http://nomisma.org
75 At the time of writing, there were 31,201 unique coin entries hosted on Nomisma. These all refer to the subject URIs minted by the host institutions.
dataset. Figure 29 shows a map of all of the known find spots and mint locations for all of the collection’s coins. This is a successful use of a map visualisation to present distribution data. It clearly highlights the geographical areas from which most of OCRE’s coins originated.

Figure 29: a distribution map showing the find spots and mint locations for the OCRE coins

In many respects Nomisma is a model Linked Open Data resource for the cultural heritage sector. Besides the digital databases associated directly with the activities of the American Numismatic Society, Nomisma has been used to provide thesauri for many
other institutions around the world, such as the British Museum, the UK’s Portable Antiquities Scheme and more recently the University College Dublin Classics Museum. At a technical level, before Nomisma made the decision to move to a Linked Open Data platform, it was built upon a XHTML content base, which used Apache Tomcat, Cocoon, Solr, Orbeon and eXist. Its Linked Open Data services are now provided by an Apache Jena RDF triplestore system. Fuseki is used to provide a SPARQL interface to the RDF data and all of its URIs can be retrieved as RDF data in various serialisations.

Nomisma’s data is structured using a custom ontology, which is available on its website. The ontology is largely flat of hierarchy, which makes it relatively straightforward to implement (Figure 30). Nomisma also employs the use of Dublin Core Terms and the Basic Geo Vocabulary (Brickley 2003). The success of Nomisma is no doubt helped by the fact that the study and taxonomy of Roman Imperial coins in particular is very well established. For instance, the compilation of the multi-volume series Roman Imperial Coins catalogue began in 1923. This research environment has created a high level of taxonomic standardisation within the field and because of this, ancient coins and their types are ideal candidates for representation using RDF.

**PeriodO**

Dates and periods have always presented archaeologists with one of their sternest ontological challenges (Binding 2010; Pare 2008). Even in the pre-digital age, being able to chronologically interrogate disparate information sets, each using its particular periodization system, has presented enormous difficulties. How might one conceptualise the relationship between Artefact A, which has been given the label ‘Late Iron Age’ with another labelled ‘Early Archaic’? Is there an overlap in time between the two subjects? How long do these periods last? Might one artefact have been in use for just a short period, perhaps at the beginning of a very long block of time, while the other enjoyed a much longer period of use? The problem becomes even more entrenched when one considers that periods are often spatially contingent – for example, the ‘Iron Age’ in the eastern Mediterranean might well represent an entirely different range of calendar years to that of the Iron Age in north-western Europe.

---

76 [http://nomisma.org/ontology](http://nomisma.org/ontology)
Artefacts serve as invaluable proxies for archaeologists. Their analysis and the dates given to them allow archaeologists to date the stratigraphy within which they were found and in turn date the site that contained that stratigraphy (Renfrew & Bahn 2004 pp. 122–123). Because of this, material dating has always been recognised as one of the key tools at the archaeologist’s disposal and issues such as site contemporaneity (Dewar 1991) and the re-dating of multiple sites as and when new discoveries are brought to light, ultimately emanate from what are often necessarily fuzzy artifactual timescales. This temporal ‘fuzziness’ is exacerbated in the case of prehistoric archaeology.

The PeriodO project77 (Rabinowitz et al. 2015) was established to deal with this very problem using Linked Open Data techniques. As Rabinowitz notes, periods are ‘essentially arbitrary conventions about which scholars disagree’ (Rabinowitz 2014 p. 1). However, despite this perceived arbitrariness, periods form the basis of the archaeological method. Rabinowitz et al. have realised, quite rightly, that the period problem is an ideal candidate for resolution using Linked Open Data methods. The ability of Linked Open Data datasets to be dynamic as a result of their being connected to other (possibly changing) datasets means that the variability of periods, which in the pre-digital world had demanded such enormous intermittent reinvestments of labour, now becomes an attractive feature of a dynamic linked knowledge system.

PeriodO is, essentially, a gazetteer of period values that are created by its community of users. Rabinowitz is at pains to point out that the knowledge production for the system will follow a ‘bottom up’ model, eschewing the ‘top down’ knowledge that characterised so much of pre-digital archaeological practice. Democracy of opinion fits well with the Linked Open Data philosophy. However, despite this potential for a more liberal knowledge environment, for the most part, the majority of Linked Open Data creation has originated in the corridors of the larger knowledge institutions and, as such, PeriodO’s attempt to swim against this general current should be applauded.

77 http://perio.do
Technically, PeriodO presents what it calls ‘period assertions’ as concepts that are modelled using SKOS. As such, PeriodO’s data is a vocabulary of terms that are exposed as subject URIs and which include links to citation and contextual information about the period in question. As is shown in Figure 31, the ‘Iron Age’ in the Levant has been understood in a number of different ways over the years by scholars with Aharoni assigning it a date range of ‘1200-586 BC’, while Younker is less specific in 2003 as he gives it a more descriptive date range of ‘1200 BC – mid-6th century BC’. Others, as would be expected, have slightly different understandings of the Iron Age in Israel. These are all valid readings of this period in their own way. Consumers of PeriodO data will have access to all of these various interpretations; with no one source gaining precedence over another.

It is the intention of the project to serialise its RDF data as JSON-LD and its DOIs are to be minted using the EZID system of the California Digital Library. The dataset will be hosted on Github and it is hoped that the project will eventually also provide a SPARQL endpoint for its data.

---

78 See Chapter 3 for an overview of where SKOS fits within the Semantic Web model.
79 See Chapter 2 for a description of all of the various RDF serialisations currently available.
Pleiades and Pelagios

The Pleiades and Pelagios projects have much in common with the Nomisma and PeriodO initiatives both in terms of the technical strategies that they have employed and in their basic objectives. Whereas PeriodO takes as its subject matter the vocabularies that archaeologists use to assign periods to material and sites and Nomisma is interested in the concepts employed by the numismatic community, Pleiades looks to ancient place-names for its subject matter (Simon et al. 2012).

Place-names change through time and vary across languages. How might a human user or, more significantly, a computer realise that the terms Athens, Athenae and Αθήνα all refer to the same conceptual or geospatial entity? This potential for misinterpretation becomes a huge problem when you are dealing with multiple sources, which might have originated in very different contexts, be they temporal or spatial variations. In the example just cited, Athenae was the term used between 750 BCE-640 CE to describe what is now the city of Athens in English or Αθήνα in Modern Greek. For a human, this collection of terms to describe Athens might not prove an insurmountable challenge. However, the advanced Natural Language Processing techniques that humans take for granted and, which allow for the handling of such processing with apparent ease, would present a sizeable problem for a computer and, as we have already seen, it is intended that Linked Open Data resources be consumed primarily by computers.

The Pleiades service aims to solve these place name ambiguities by providing a URI for each and every spatial concept in its datastore. It was originally designed to deal exclusively with ancient place names but more latterly, it has begun to spread its net wider across more modern datasets (Simon et al. 2014). The Pleiades website can be joined by any member of the public – a user need not have any particular academic credentials. Any user is entitled to create new Pleiades data although this does not become an official part of the Pleiades listing until it has been vetted by the community of peer reviewers.

---

80 http://pleiades.stoa.org
81 http://pelagios-project.blogspot.com
82 Ancient meaning ‘antique’ in this case.
Each Pleiades entry is ontologically structured as follows (Figure 32). A pleiades:Place entity links to pleiades:Location entity via the pleiades:hasLocation predicate. The pleiades:Place entity contains documentary data such as the name, type, description and creator of the resource. It also contains links to bibliographic resources that substantiate the claims contained within the other associated data. The pleiades:Location entity details the geospatial coordinates of the place in question – it can be a point or a polygon. pleiades:Location also expects the user to specify a start and an end date that associates this particular location with the relevant pleiades:Place entity for that period of time. The logic here is that, while a geospatial specification is essentially eternal, places are transient and come and go out of existence with the passage of time.

![Figure 32: the Pleiades data model](image)

Unsurprisingly, Pleiades data is structured as RDF triples. A Pleiades namespace exists that exposes an ontology of classes, predicates and vocabulary values and these are used alongside the ever-present FOAF, Dublin Core Terms, SKOS and Geo ontologies. Some less well known ontologies used by the project are the PROV Ontology (Belhajjame et al. 2013) and the Citation Typing Ontology (Shotton 2010).

The Pelagios project builds on the foundations of Pleiades. Its goals are twofold: to make it easier for data content providers working on material related to the ancient world to
publish place names and to make it easier for users interested in consuming ancient place name data to do so across all of the datasets that expose this type of information (Simon et al. 2012).

Pelagios is a community of users who contribute content that they have curated. The idea is that by creating an aggregation of all of the Linked Open Data datastores that contain references to ancient place names, users will be able to find links between these datasets based on their inclusion of canonical Pleiades place name URIs. Besides this requirement to use Pleiades URIs, partners must also model their place name data using the Open Annotation (Sanderson et al. 2013) and Vocabulary of InterLinked Open Datasets (VoID) (Alexander et al. 2011) ontologies.

**The Getty Linked Open Data resources**

The Getty vocabularies were first formalised in the 1980s and their aim then and now is to ‘help people categorize, describe, and index cultural heritage objects and information’ (Alexiev et al. 2014). There are four Getty thesauri (AAT, TGN, CONA and ULAN), all of which are compliant with ISO and NISO standards for thesaurus construction, although to date only the first two have been published as Linked Open Data resources.\(^{84}\)

The Art and Architecture Thesaurus (AAT) exposes generic concepts related to the fields of art, architecture, broader cultural heritage contexts and conservation. AAT concepts can be used to describe the basic type of a particular work of art or architecture, its stylistic conventions, its component material parts and the subject matter that it deals with. The AAT is concept-centric and it is polyhierarchical, in the sense that any one concept can have more than one parent. The AAT as of August 2014 held 42,000 concepts.

The Thesaurus of Geographic Names (TGN) is concerned with places and their names. The TGN essentially tackles the same problem as Pleiades; places can be associated with multiple names and these can change through time. The TGN contained 1.26 million places and 1.85 million names linked to these place concepts in August 2014.

---

\(^{83}\) http://www.getty.edu/research/tools/vocabularies

\(^{84}\) This was correct as of 20 March 2015. The ULAN thesaurus was scheduled to be made available as Linked Open Data in April 2015 and the CONA by the end of 2015.
Figure 33: Arthur Evans as represented by the Getty ULAN data service

The Union List of Artist Names (ULAN) vocabulary was created by Getty as a repository of names chiefly associated with practitioners of the arts, although as we will see, this list is not exclusively populated with artists. For example, the ULAN subject 500212319 groups information about the British archaeologist, Arthur Evans (Figure 33). The ULAN record includes fields such as the various names, nationality, roles (for example archaeologist, et cetera), gender, birth and death dates, and citations related to Evans. 582,000 names were represented by the ULAN in 2014.

The Cultural Objects Name Authority (CONA) dataset contains information about specific fixed and movable works of art and architecture. The CONA is currently only available to view online in a limited form. However, the few records that do exist exemplify the agenda of the project quite well. Take for example the record with the subject ID of 700000206. This details information about the Pergamon Altar, which is currently housed in the Collection of Classical Antiquities in the Berlin Museum. The CONA record includes information about the altar's various titles, its type (which points
to the AAT altar concept), its classification as a piece of architecture, its creation date in the 2nd century BCE, its current and past locations, style (another AAT reference), related works, subject matter, citations and general notes. CONA is, therefore, a consumer of the AAT, TGN and ULAN. While the three vocabularies describe idealised concepts, the CONA’s data represents material objects.

The Getty vocabularies present a vast range of concepts that can be used by cultural heritage professionals to categorise their material. They allow for the categorisation of most conceivable aspects of cultural heritage material objects. And now with their provision of Linked Open Data interfaces (Harpring 2014), they are making it possible not only for big institutional museums and galleries but also for smaller initiatives, such as the kerameikos.org project,85 to utilise their resources.

The Getty concepts have been constructed and arranged over a number of decades through a process of consultation with the GLAM (galleries, libraries, archives, and museums) sector (J. Paul Getty Trust 2014). The sifting through of these contributions is organised chiefly by Patricia Harpring of the Getty and the technical requirements are managed by Vladimir Alexiev and Joan Cobb.

Technically, the Getty vocabularies are built using the SKOS ontological model. SKOS, as was discussed in Chapter 3, has become for all intents and purposes the de facto standard for the structuring of concepts within RDF datasets, particularly controlled vocabularies. Getty vocabularies also utilise the Dublin Core (DC) and Dublin Core Terms (DCT) ontologies to represent common properties of the concepts that they model. The Bibliographic Ontology (BIBO) and Friend of a Friend (FOAF) are used to render related source materials and contributors. Version control is an important aspect of the Getty Linked Open Data system and it is a data curation feature that is largely ignored by the other projects covered by this review. The Provenance (PROV) ontology is used to provide properties related to this class of information.

85 http://kerameikos.org provides a Linked Open Data resource for Greek pottery types and objects.
Figure 34: the semantics of the Getty Vocabulary Program ontology
The Getty data also makes extensive use of its own in-house Getty Vocabulary Program (GVP) ontology (Figure 34). The GVP is an extension of the SKOS ontology with the gvp:Concept class taking centre stage following the example set by the skos:Concept class. The polyhierarchical structure of the AAT and the TGN is provided for using a series of predicates (gvp:broader, gvp:narrower, gvp:broaderExtended, gvp:narrowerExtended, gvp:broaderPreferred, gvp:broaderPreferredExtended, gvp:broaderNonPreferred, gvp:broaderGeneric, gvp:broaderPartitive, gvp:broaderInstantial, gvp:broaderGenericExtended, gvp:broaderPartitiveExtended, gvp:broaderInstantialExtended), which expand upon the SKOS predicates that deal with relationships between concepts.

![Image](image_url)  
**Figure 35:** An HTML view of the Getty Kamares concept (Harpring et al. 2015)

The AAT and TGN records can be accessed as various RDF formats such as JSON, RDF Turtle, RDF XML and this is implemented simply and well (Figure 35). The data is also made available via a SPARQL endpoint and while this is encouraging it appears to be at an early stage of its implementation and often delivers poor performance and reliability.\(^{86}\)

\(^{86}\) This observation is made on the basis of the author’s experience of using the system on 23 April 2015.
Europeana

Europeana (Europeana Project 2015a) is another data aggregator that makes particular use of Linked Open Data technologies to organise, link and disseminate its data. Launched as a prototype in 2008, it is intended to act as a portal into the digital resources of all European libraries and cultural heritage archives. As of 23 April 2015, it holds 39,276,880 digital objects (Europeana Project 2015b), of which 49% are made available for re-use under various licenses. 23,213,500 of these digital objects are related to images, 15,129,531 to text documents, 507,463 to sound resources, 408,553 to video resources and 17,833 to 3D data. Users are empowered to browse by theme, by faceted search or to construct complex SPARQL queries to interrogate the dataset.

![Europeana Digital Object](image)

**Figure 36: a view of the Europeana digital object related to the site of Newgrange, Ireland (Europeana Project 2015c)**

Europeana itself is not a repository for the raw data associated with these resources. Instead, its data details the source object’s context or metadata. The raw data for the object can be located on the source institution’s website, if required. For example, the Europeana digital object representing the site of Newgrange in Ireland lists a title, description, subject, identifier, language, rights, and country field (Figure 36). It also lists the URL for the provider who submitted the content to Europeana.

All Europeana content is created by its content partners, who typically come from the museum, gallery and archive sectors of European member states. Europeana sets out

87 http://www.europeana.eu

88 A Europeana object is defined as ‘a digital representation of an object that is part of Europe’s cultural and/or scientific heritage’ (Fallon 2015)
guidelines as to the form of this data and the onus is on the source institution to inject as much context into the data in the form of metadata as is possible (Pekel 2015). They are, however, just guidelines and there are few actual baseline requirements stipulated by Europeana. This metadata is then sent on to the Europeana import teams, who carry out the processing required to transform the structure of the data, before it is finally imported into the Europeana datastore. At the time of writing, a dataset submitted to Europeana will appear as publically available live data within about three months.

Europeana data is structured using the Europeana Data Model (EDM) (Europeana Project 2014; Isaac 2013). The EDM is an evolution of and is backward compatible with the Europeana Semantic Elements (ESE) ontology (V. Charles 2015), which was created during the first iteration of the project. As we will see, the EDM is semantically fairly rigid but it does allow for flexibility in the form of thesauri that are used to populate its fields (Peroni et al. 2013).

The EDM draws a distinction between the real-world object and its digital representation. The material object is modelled as an instance of the edm:ProvidedCHO class and the digital representations associated with this object are modelled using the edm:WebResource class. The combination of the single edm:ProvidedCHO class instance and the potentially infinite edm:WebResource class instances are modelled using the edm:Aggregation class and each edm:Aggregation class is coupled with a edm:Proxy class, which allows for the possibility that two institutions might submit data about the same real-world object.

The EDM also provides classes for the modelling of what are described as ‘contextual’ entities. These are concepts that flesh out the real-world contexts or subject matters of the objects in question. For example, the edm:Agent class can be used to model people or organizations that appear as subjects in a painting. Similarly, the edm:Event, edm:Place, edm:Timespan classes allow Europeana data providers to encode event, place and time meaning for the digital objects that they deliver. The EDM makes use of the SKOS:Concept class to inject controlled vocabulary meaning into objects and their representations. The inclusion of these EDM and SKOS classes is intended to allow the

---

89 The consistency of these ontological demarcations has been questioned by some scholars (Peroni et al. 2013)
user to ask the questions, ‘Who?’, ‘What?’, ‘Where?’ and ‘When?’ of any one of the digital objects that it hosts.

![Diagram of an object-centric modelling of Da Vinci's Mona Lisa by Europeana](Isaac 2013 fig. 7)

Europeana data providers can choose either to model their data using an object-centric or an event-centric model. The first model puts the material object at the centre of the conceptual mapping. Properties are used to flesh out the characteristics of that object and these tend to be rendered using properties taken from the Dublin Core metadata scheme (Figure 37).

![Diagram of an event-centric modelling of Da Vinci's Mona Lisa by Europeana](Isaac 2013 fig. 10)

The event-centric model (Figure 38), with its foregrounding of the human role in the assignment of meaning to a target object, has already been discussed in great detail in the context of the CIDOC CRM model in Chapter 3. As with CRM data, applying Europeana’s event-centric approach tends to produce more data than one centred on the
object but the character of this data can be much more nuanced than that of object-centric conceptual systems, allowing for far greater flexibility of interpretations.

Figure 39: a graph view of the Europeana Newgrange resource

As with the other Linked Open Data resources that are discussed in this section, Europeana data can be consumed by the user in a number of different formats: HTML, JSON, RDF/XML, RDF Turtle and RDF N-Triples. The HTML view also allows the user to view the data as a colour-coded graph (Figure 39), which greatly eases the conceptual understanding of the data’s underlying structure. The data is also available via a SPARQL endpoint, as one would now expect of any serious Linked Open Data resource.

While there is much to laud in the Europeana project, both in terms of its stated objectives and in the manner of its delivery of targets to date, there are certain deficiencies that should be noted. First, the EDM vocabularies make extensive use of string values and not the URIs that one would expect of a Semantic Web resource. For example, values linked to subjects via the edm:type predicate come from a pool of the
following string values: ‘IMAGE’, ‘TEXT’, ‘SOUND’, ‘VIDEO’ or ‘3D’. The problem with this approach, is that incorrect spellings will be accepted by the system and yet if an RDF resource has an edm:type of ‘IMGE’, it will be missed by queries that are looking for resources with edm:type of ‘IMAGE’.

Peroni et al. (Peroni et al. 2013) also refer to the fact that the vast complexity of the Europeana dataset allows for situations within which the same real world object can come to be rendered inconsistently by different data providers. The example they highlight is that of Giambattista Piranesi’s publication ‘Le antichità Romane’. One data provider might decide to define the object to be of type ‘print’, while another classifies it as type ‘amphitheatre’. As in the case of the incorrect spelling of the ‘IMAGE’ resource cited above, this type of inconsistent labelling of digital resources makes it more difficult for consumers to ultimately find those resources. Of course, the other side of the argument is that it is important for Europeana to represent the full spectrum of views of its provider community. How to reconcile these two positions is one of the great challenges that confront projects such as Europeana.

Open Context

The Open Context project has been delivering information system services to the archaeological community since 2007 (E. C. Kansa 2012; E. C. Kansa & Kansa 2011; S. W. Kansa et al. 2007; S. W. Kansa & Kansa 2007). The Alexandria Archive Institute (AAI) developed Open Context to address the central problem outlined by this thesis; that of the linking and aggregation of disparate cultural heritage datasets (S. W. Kansa et al. 2007 p. 189). The particular model of data sharing used by Open Context is ‘data sharing as publication’ (E. C. Kansa 2014), in which the creators of the data (the ‘authors’) work with a team of specialised data ‘editors’ towards that data’s publication.

There is a lot to be said in favour of this approach as it is clear that the practicalities of publishing data as Linked Open Data assume a specialised technical knowledge set and demand a significant effort on the part of the data creator and such knowledge and effort would be more efficiently delivered by teams of experts in that field. Conversely, it also

---

90 This information was retrieved from the Europeana SPARQL endpoint on 7 April 2015 using the following SPARQL query:

```
SELECT DISTINCT ?type WHERE { ?s edm:type ?type }
```

91 http://opencontext.org
makes sense to have the data ‘authors’ involved at all levels of its publication, as no one
knows a dataset’s character and particularities better than its author.

At the time of writing, Open Context has published the data of 54 cultural heritage
projects (Figure 40), the majority of which being archaeologically themed with a
particular emphasis on sites found in the general area of the Near East. Each project is
divided into a number of tables that contain data related to that table type. Examples of
table types are find types (coin, glass, animal bone, human bone, pottery), archaeological
site organisational units (field project, trench, lot, feature, locus, context) and
geographical entities (region, area, site). The records that populate these various tables
tend to be interlinked with records found in other tables. For example, a lot record might
contain links to records in an animal bone table.

At the interface level, Open Context appears to the user as one has come to expect of
any online information system (Figure 41) and indeed its design is not too dissimilar to
that of the linkedarc.net web app, as we will see. Text and image data is presented to the
user as cells within tables. Data is geotagged and a map appears as a sidebar giving the
record its geographical context. Following the Open Context publication analogy, all data
records are given a rating for their editorial status, with 5 stars being awarded to data that
has been fully peer-reviewed, while poorly documented data is given only a single
editorial star. Any references associated with a record are displayed prominently as
categorised lists. Open Context contains a flexible, speedy and easily accessible search interface. The results of a free-text search are categorised variously into context, category, descriptive properties and related people facets and these facets can be selected to narrow down the search result set.

Technically, the Open Context system has seen a number of design iterations over the years. These have come about as a function of changing technological possibilities but also as the objectives of the project have evolved. The original system was built using a MySQL, PHP and Dojo AJAX stack. It employed as a data ontology the services of the Archaeological Markup Language (ArchaeoML) data model, which was developed by David Schloen of the University of Chicago.\(^2\)

In recent years, the project has begun to move in the direction of Linked Open Data and the Semantic Web and its technological stack has evolved to reflect this. To begin with, Open Context introduced a number of RESTful web services that, while proprietary, have at least heralded a new period of openness for the project. More recently, this transformation has accelerated with the deprecation of the ArchaeoML standard and the implementation of the CIDOC CRM on a test case basis alongside an RDF data representation. The project has also begun to use Pleiades URIs to deliver geotagging for its data.

---

\(^2\) See Chapter 2 for more on ArchaeoML.
The Linked Open Data experience or more specifically the decision to use open ontologies has, however, not been an entirely positive one for the project and Kansa (2014) outlines how the effort involved in the implementation of a highly complex conceptual model such as the CRM may not be equal to its potential dividends. By way of comparison, Kansa notes how time spent on a parallel zooarchaeology-themed project, mainly on the honing and implementation of the Encyclopaedia of Life (EOL) controlled vocabulary to be used on zooarchaeological material, delivered more rewards in terms of tangible benefits for Open Context’s user base. The lack of easily accessible examples that describe best practice to potential CRM adopters is one major problem identified by Kansa and he also cites how even such an ordered knowledge information system as the CRM cannot avoid scenarios in which guidelines are interpreted differently depending on the context and how in these cases data can come to populate a system inconsistently.\(^\text{93}\)

**British Museum**

The history of the British Museum’s embrace of Linked Open Data and Semantic Web technologies is very much a history of an extensive implementation of the CIDOC CRM model. The museum’s relationship with the CRM began in 2010 when it received a Mellon Foundation grant to support the ‘planning of a collaborative online research environment for art historians, curators, conservators, and scientists’ (Mellon Foundation 2015). The museum’s exploits in the digital world, however, can be traced back further to 2007 when Collection Online, its online database was first launched (British Museum 2011).

The institution’s Semantic Web investigations have been conducted largely under the auspices of the ResearchSpace initiative (Oldman et al. 2015) and to date the outputs of the project have been extremely impressive. The museum was the first UK arts institution to engage seriously with Semantic Web approaches to data curation and delivery (British Museum 2011) and this pioneering status is now evident in the maturity of their Linked Open Data offering. Given the size of the museum’s collection, it is a

\(^{93}\) Kansa uses the example of a Munsell entry to illustrate his point about inconsistent data entry. Data inputter A chooses to record the reading for a particular artifact using the P3_has_note CRM predicate, while inputter B interprets the reading to be a measurement, which should be associated with the artifact using the P43F_has_dimension predicate instead. Both of these readings can be argued to be valid implementations of the CRM.
testament to the professionalism of the ResearchSpace project that so much of its material is now being made publically available as digital data in such an extensive and consistently structured manner.

Figure 42 shows a view of a record for a 26th Dynasty sarcophagus that was retrieved using the British Museum’s SPARQL interface. A small thumbnail image of the resource is displayed in the upper left corner of the page. This is followed by a table of predicate and object pairs that are linked to the sarcophagus subject.

While the sarcophagus resource is structured using other ontologies (the British Museum’s own bmo ontology chiefly), the CIDOC CRM’s ecrm prefix dominates the predicates list. This type of Linked Open Data interface should now be familiar as its design and the user experience that it delivers is consistent with the approach taken by the majority of the other Linked Open Data content providers discussed in this section. The user is able to click on predicates or objects that are composed of URI values. Clicking on these URIs navigates the user to a new page, which in turn displays a summary of the predicate-object pairs for the new subject URI.

The user can also choose to filter the predicate-object return sets by named graph and/or language. He or she might also decide to select how they would like the SPARQL engine

94 http://collection.britishmuseum.org/sparql
to perform its inference processing. As was discussed in Chapter 3, drawing inferences between graph data nodes is a powerful tool in the hands of the Semantic Web consumer, as it allows search criteria to be expanded on the basis of the inference rules that are coded by the underlying ontologies and this can lead to unexpected and rewarding outcomes.

For example, if we asked the SPARQL endpoint to return all of the subjects that are of type E39_Actor and the inference engine were to be turned off, the SPARQL engine would return only those subjects that are explicitly of this type. E39_Actor is, however, the parent class of E21_Person and E74_Group and so the user might legitimately be interested in viewing subjects of these types as well. This approach to data interrogation is familiar to humans, as it is the way that humans brains operate when faced with taxonomic problems (Hotte & Muller 2006). For computers, however, drawing inferences, while equally powerful, comes with a processing overhead and this needs to be taken into account when considering its use as part of a Linked Open Data solution.

**Seneschal**

Free text object values are the enemy of the Linked Open Data implementation because they introduce the potential for data variance and error, which makes the data less indexable and ultimately less valuable as a result. Free text fields are constrained only by the alphabet that is used to complete them – data inputters are free to enter whatever text strings they wish into such fields. At one level, this makes the field highly flexible, which might be favourable in certain circumstances but with great flexibility also come disadvantages.

Take for example, a set of data that describes the artefacts that are uncovered and recorded at an archaeological site. The creator of the artefacts’ digital records will most likely need to include a type field that defines the type of any one object. If the field is determined to be a free text field, the data inputter is free to enter whichever text they feel describes a feature best. So for a Roman lamp object, they might enter the text ‘lamp’. If it was deemed to be a lamp of diminutive dimensions, they might instead prefer to enter ‘small lamp’ into the field. The free text field will accept any of these values as valid data. The problem becomes exacerbated when a second individual starts contributing to the same dataset. They might choose to define a second lamp artefact as a ‘Late Roman lamp’. If English was not their native language, they might instead enter
the French for lamp, ‘lampe’, or perhaps the Chinese 灯. And how is one to deal with word ambiguities? What might the English word ‘wreck’ refer to? Is it a shipwreck? A wreck of a car? And so on.

The Seneschal project\(^{95}\) (Binding 2014b, 2015a) targets this problem of free text use for the definition of types within cultural heritage datasets. For conceptual data such as types, it is far preferable to use URIs instead of free text because it removes the ambiguity that is outlined above. The Seneschal project has harvested the vast typological catalogues that have been compiled and used by a number of the leading UK cultural heritage bodies (English Heritage, Royal Commission of Ancient and Historic Monuments of Scotland (RCAHMS), Royal Commission of Ancient and Historic Monuments of Wales (RCAHMW)) and turned them into Linked Open Data resources. The project’s goal is to make these resources easily discoverable by creators of cultural heritage content (it need not only be Linked Open Data) and therefore to encourage their re-use within the community. This promulgation of the core set of established controlled type vocabularies will not only help the individual projects who choose to adopt them but also the wider community as data comes to be standardised across the whole user domain.

Seneschal has two corpuses of material in mind. Firstly, it has been working diligently on aligning existing datasets to the vocabularies that it hosts. This alignment work is complex and involves techniques borrowed from the world of Natural Language Processing. Essentially, algorithms are written, which crawl through tens of thousands of free text fields and decide how to map each value on to a particular set of thesaurus values. The probability that a match is correct is then calculated (Table 14) and the system automatically decides whether to accept the mapping or to flag it up for manual inspection by a human.

\(^{95}\) http://www.heritagedata.org/blog/about-heritage-data/seneschal
The second set of datasets that Seneschal targets is new cultural heritage datasets. By encouraging the creators of these datasets to access and use their services, they hope to create a future ecosystem of cultural heritage datasets created using standard vocabularies.

The data types covered by the Seneschal project are expansive and reflect the years that were involved in their creation and the range of subject matter that they apply. English Heritage has contributed the Event Type (terminology used to describe archaeological activities), Monument Type (classification of monument types), Archaeological Sciences (types related to the application of archaeological scientific methods), Building Materials (material used to construct monuments), Maritime Craft Type (ship types), Period (periods), Components (elements of a building) and Evidence (types related to evidence of archaeological remains) vocabularies.

The vocabularies from the RCAHMS and the RCAHMW overlap with the content of the English Heritage thesauri, reflecting parallel but often-independent developmental paths taken by the three groups. The RCAHMS has contributed the Monument Type Thesaurus (Scotland) (monument types found in Scotland), Archaeological Objects Thesaurus (Scotland) (artefact types found in Scotland), Maritime Craft Thesaurus (Scotland) (boat types found in Scottish waters) and from the RCAHMW Seneschal lists

<table>
<thead>
<tr>
<th>Data value</th>
<th>Highest scoring thesaurus term match</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>AXE FACOTRY</td>
<td>Axe Factory</td>
<td>90%</td>
</tr>
<tr>
<td>BOUNDARIES</td>
<td>BOUNDARY</td>
<td>77%</td>
</tr>
<tr>
<td>BOUNDARY</td>
<td>BOUNDARY</td>
<td>100%</td>
</tr>
<tr>
<td>BUIED SOIL HORIZON</td>
<td>BURIED SOIL HORIZON</td>
<td>97%</td>
</tr>
<tr>
<td>CAIRN</td>
<td>CAIRN</td>
<td>100%</td>
</tr>
<tr>
<td>CAIRN (POSSIBLE)</td>
<td>CAIRN</td>
<td>100%</td>
</tr>
<tr>
<td>CAIRNN</td>
<td>CAIRN</td>
<td>90%</td>
</tr>
<tr>
<td>CESS PIT</td>
<td>CESS PIT</td>
<td>94%</td>
</tr>
<tr>
<td>CHAMBERED TOM</td>
<td>CHAMBERED TOMB</td>
<td>96%</td>
</tr>
<tr>
<td>COMERCIAL</td>
<td>COMMERCIAL</td>
<td>94%</td>
</tr>
<tr>
<td>CROFT?</td>
<td>CROFT</td>
<td>90%</td>
</tr>
<tr>
<td>CUP-MARKED STONE</td>
<td>CUP MARKED STONE</td>
<td>93%</td>
</tr>
<tr>
<td>DICTH</td>
<td>DITCH</td>
<td>80%</td>
</tr>
<tr>
<td>ENCLSOURE</td>
<td>ENCLOSURE</td>
<td>88%</td>
</tr>
<tr>
<td>EXTRACTION PIT</td>
<td>EXTRACTIVE PIT</td>
<td>85%</td>
</tr>
<tr>
<td>EXTRACTIVE PIT</td>
<td>EXTRACTIVE PIT</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 14: matching free text data to thesaurus terms by the Seneschal project (Binding 2014b)
the Period (Wales) (periods in use in Wales) and the Monument Type Thesaurus (Wales) (Welsh monument types) sets.

At a technical level, Seneschal employs a combination of Linked Open Data strategies that are now very familiar. The data is modelled as RDF using the SKOS ontology, which is a predictable choice given the focus of the project on conceptual data. The RDF data is made available to the user via a number of web services (Binding 2015b) and a SPARQL endpoint. The Seneschal team have also made available a number of so-called widgets, which are blocks of customisable web user interface code that can be easily integrated into a user’s website (Charno 2013). The objective here is to give third party projects access to the Seneschal vocabularies at the point at which their data is being created. These projects might want to add a widget to the data entry page on their website, which will allow their users to search for and use URIs taken from Seneschal.

Figure 43: an English Heritage crop mark concept as viewed on the Seneschal system

---

96 http://heritagedata.org/live/sparql.php
Chapter 4

The Seneschal website’s user interface (Figure 43) is designed with the experienced cultural heritage data creator in mind. The core function of the site is to present its vocabulary entries as tables of properties and linked values. These are supplemented by pages that detail the technical aspects of embedding Seneschal widgets into a web app, consuming its resources using web services or using its SPARQL endpoint.

**CLAROS**

The Classical Art Research Online Services (CLAROS) project\(^7\) is an online resource of ancient artifactual material. It was established in 2000 to provide an authoritative dataset of ancient Greek and Roman material but since 2011 it has come to embrace a more liberal geographic and temporal range by welcoming data from the Islamic, Chinese and Indian art into its fold (Kurtz 2012).

Based in and heavily associated with the sphere of digital humanities research at the University of Oxford, it obtains its data from content partners who include such cultural heritage luminaries as the Ashmolean Museum, the Bodleian Museum and the Pitt-Rivers Museum (Kurtz 2012). CLAROS hosts many types of data, which are divided into three broad subject groups: objects, people and places. And as with the Europeana project, these categories allow the user to ask the questions ‘who?’, ‘what?’, ‘when?’ and ‘where?’ of the dataset.

CLAROS has been employing Semantic Web technologies since 2007 and, as such, can be viewed as an early adopter and of particular interest to our review. As with the British Museum, it chose to use the CIDOC CRM to model its data. However, while the British Museum can be seen to have applied a particularly expansive implementation of the model, CLAROS has taken a more pragmatic and minimalist road, choosing instead to model only those aspects of its dataset deemed to be of most value to its consumer base (CLAROS 2015). In that light, CLAROS data subjects tend to link to a limited but easily understandable set of properties, such as title, provenance, associated visual representation, type, condition, source material and current location (Figure 44).

---

\(^7\) [http://www.clarosnet.org](http://www.clarosnet.org)
Figure 44: a coin record displayed on the CLAROS system

CLAROS provides access to its data via custom web services and a SPARQL endpoint. Helpfully, the web app includes a listing of the SPARQL query used to generate the raw data for the current page and in that regard CLAROS has clearly been designed with the writer of third party cultural heritage applications in mind.

Linked Open Data archaeological resource patterns

The preceding review shows that quite a number of archaeological and cultural heritage groups are now hosting data services that are consistent with Linked Open Data practices. The reasons for this are myriad and complex and we will consider these contexts in some detail in Chapter 7 but above all else it is clear that Linked Open Data practice sits well within an archaeological discipline that values the free flow of knowledge.

It is possible to divide the preceding list of archaeological Linked Open Data service providers into three broad categories (Figure 45). In the first category are the concept providers. This group provide taxonomies of particular classes of information, which can then be used by Archaeological Information Systems to describe their digital objects. The second group, which we will call the content aggregators, bring together disparate cultural heritage datasets by aggregating their content via a unified Linked Open Data interface. Finally, there are the individual collection data providers and while these share
many of the characteristics of the second group, they are solely concerned with the representation of a single set of data.

The majority of the concept providers apply very similar methodological mixes (minted DOI provision, RDF, SKOS modelling, SPARQL and web service interfacing) but they do so often with very different subject matters in mind. Pleiades deals solely with geographical entities, Nomisma numismatic material, and PeriodO the boundaries that archaeologists use to delinate time. Getty and Seneschal have a wider remit, however, as they both deliver concepts that cover all of these conceptual groupings and more (architectural and artifactual elements, ship types, et cetera).

The emergence of these five concept providers has heralded the beginning of a new era of knowledge sharing within archaeology. Archaeologists now have what amounts to a small but growing marketplace of Linked Open Data conceptual resources that are free to use and built upon a solid platform of community support. Tellingly, they are services that are now being actively used by content providers working in this space. For example, Nomisma DOIs are now being used by the British Museum, by German
numismatic collections and more recently by the Classical Museum at University College Dublin. Pleiades DOIs have proved hugely popular and have seen adoption by groups such as the Portable Antiquities Scheme in the UK and FastiOnline. The Getty has a long history of international adoption within the museum sector in the pre-digital world and with the launch of its catalogues as Linked Open Data, this will no doubt transfer over to the digital world as museum collections slowly make strides towards digital practice. The team behind Seneschal are similarly building upon an impressive lineage of scholarship and given the lengths that they have gone to make their resource accessible by the current and future cultural heritage digital data providers, their DOIs should also begin to appear among the many digital collections that will benefit from the use of more standardised typologies.

The Linked Open Data cultural heritage aggregator projects have a difficult task, as they need to provide a homogeneous interface to what are typically heterogeneous datasets. Kansa’s model of the data aggregator as ‘editor’ of data that an ‘author’ submits for review and publication is very useful in this regard. The publishing industry is built upon the premise that editors do the work of moulding an individual’s written piece into a standardised format that eventually appears alongside other similarly moulded works. While this approach can deliver numerous benefits in terms of the scale of data that is made available to the user, it also comes with disadvantages as standardisation necessarily demands a degree of compromise in terms of data individuality and character.

The British Museum is the only project reviewed that falls into our last grouping of Linked Open Data providers active in the archaeological/cultural heritage space but they are not alone. The larger cultural heritage institutions would appear to want to maintain control over a lot of their own digital curation and are as a result less prepared to outsource their needs to an external group such as a content aggregator or a concept provider. The tendency of the members of this group is to look within for existing resources such as controlled vocabularies before looking to see if these resources might

98 http://finds.org.uk
99 http://www.fastionline.org
100 For instance, the National Museum of Finland have worked very closely with the Linked Open Data Finland group (Hyvönen 2015) to produce a Linked Open Data representation of their dataset. The Smithsonian American Art Museum is another example (Szekely et al. 2013).
101 This does not mean to say that outsourcing is never used by the large cultural heritage institutions. For example, the British Museum and Ireland’s Discovery Programme are now using Sketchfab as a host for their 3D content output.
be provided externally. This is a natural response as the British Museum and other similar institutions have considerable research histories that have created vast institutional resources. However, in order to achieve full Linked Open Data compatibility as outlined by Berners-Lee, these groups will need to embrace a more collaborative digital future (D’Andrea 2012).

What might appear at first glance to be a sudden revolution in archaeologically themed Linked Open Data resources is perhaps better viewed as a gradual evolution of practice. Building Linked Open Data resources is a complex matter. It involves the investment of time and effort but also and more significantly it demands a shift in mind-set towards more liberal approaches to knowledge construction and engagement and this latter social transformation often proves to be the most challenging.

**linkedarc.net**

linkedarc.net is, first and foremost, an Archaeological Information System. Its primary role is to manage the archaeological information of its partner projects, and, this makes it a content aggregator in the archaeological Linked Open Data service provider typology outlined in the previous section. As a secondary consideration, but of great significance to this thesis, linkedarc.net implements the Linked Open Data knowledge philosophy. This means that any data that it hosts becomes a part of the Archaeological Semantic Web and is available for access as RDF.

The linkedarc.net system has been developed in partnership with the Priniatikos Pyrgos project, an archaeological field initiative operating in East Crete. As such, its primary intended user group is the research archaeologist. While being a part of a wider group of individuals, this user’s working practice can be characterised as being most like that of the lone scholar. While this might sound somewhat contradictory, the geographically dispersed nature of much of the work that takes place within archaeological research projects added to the fact that the discipline is broken down into sub units of highly specialised labour means that what we would typically understand to be truly collaborative work is in fact underrepresented in many sectors of archaeological research (excepting, of course, the actual practice of excavation and survey, which are highly collaborative). This modus operandi, as we will see in the coming two chapters, has

---

102 This relationship will be described in some detail in Chapter 5.
determined much of the design of the linkedarc.net system, particularly that of its user interface elements.

The linkedarc.net system has gone through three years of development. Within that time its objectives, underlying technological nature and superficial form have changed as and when the objectives of the project have been amended or as its baseline technologies have evolved. The following section presents the evolution of the project as a narrative involving numerous false starts, incidents of erred thinking and the occasional moment of inspiration.

**linkedarc.net version 1 – native RDF Turtle file system**

![Diagram of linkedarc.net version 1 server framework](image)

Figure 46 graphically summarises the first iteration of the linkedarc.net system. It is helpful to think of this as a network of interacting agents that communicate with one another through various means that will be explained as we proceed. The lines that link the various sub-systems on the plan represent this flow of information from one sub-system to the next. The network is divided up into two broad sections with an example client inhabiting the left half of the diagram and the core server components residing on the right.

This first version of the linkedarc.net system used an underlying server architecture that is found within many server setups of this type. The combination of an Ubuntu ver12.04 operating system and an Apache ver2.2.22 web server allows for the serving of web resources (documents, images, data, et cetera) to clients making requests using the HTTP
The web.py library (Swartz 2013) was used as a mechanism to deliver web services for the project and mod_wsgi (Dumpleton & Rees 2013), an extension to Apache, allowed for the serving of dynamic web pages. The decision was made to use the RDFLib (Summers et al. 2013) Python library to enable interfacing with the RDF data.

Figure 47: the linkedarc.net data hierarchy

Conceptually, the linkedarc.net server data is structured using the hierarchy shown in Figure 47. This might seem a bit confusing given that we know from Chapter 2 that RDF creates a graph of data and not the table-based structure illustrated in this figure. In RDF, any one data node can act as a parent of any other node. The apparent inconsistency between RDF’s network form and the hierarchy shown in Figure 47 is a product of the Priniatikos Pyrgos Project’s Archaeological Information System history. As this dataset was the test-bed for the linkedarc.net project, its form proved enormously influential in the linkedarc.net system’s design. It is worth remembering that the representational form of a dataset need not be the same as its higher-level ontological structure. As we describe the mapping of the Priniatikos Pyrgos data from its original FileMaker form into RDF in Chapter 5, this interplay should become clearer.

In any case, at the top of the conceptual hierarchy shown in Figure 47 is the linkedarc.net server system. It is the parent of all linkedarc.net projects, which are in turn the parents of datatables and a datatable is the parent of a linkedarc.net record. While this basic

---

https://github.com/RDFLib
structure has remained consistent for all system iterations, the data storage has changed. For linkedarc.net ver1, this hierarchy was serialised using the Ubuntu file system. Projects and datatables were represented as directories on the server and the records were serialised as RDF Turtle files, which have been described in detail in Chapter 2.

The system operated using the following set of assumptions and basic request-response logic:

1. A client makes a HTTP request to access either a linkedarc.net web service or a linkedarc.net URI resource.
2. The linkedarc.net server code determines which RDF Turtle files need to be accessed in order to answer the request.
3. These files are opened and they are fed into the RDFLib library, which parses them and makes them queryable.
4. The client request is then interpreted and the files are queried using RDFLib.
5. The response is channelled back in the requested format to the client.

While linkedarc.net ver1’s functioning logic was relatively sound, the system soon ran into performance-related issues. Collections of RDF triples that shared the same subject were stored in a single RDF Turtle file. When a client requested information about a single record, the system was able to respond quickly to the request, as it was necessary to open and parse only a single file, which did not place an undue load on the system. However, when a client requested information that could only be obtained by querying multiple records, the response time was drastically affected. At the extreme end of the spectrum, a client request that necessitated the opening and parsing of a thousand or more record files might have taken 30 seconds to send a response back to the client and given that servers are now required to deal with hundreds or even thousands of simultaneous requests, this delay was unworkable and, therefore, ultimately unacceptable.

On the basis of this performance issue, it was decided to look for an alternative to the RDFLib/RDF Turtle file datastore solution.
linkedarc.net version 2 – a hybrid triplestore using a MySQL backend

The second iteration of the linkedarc.net server replaced the RDF Turtle files with a MySQL data backend (Figure 48). MySQL has a long and established pedigree in delivering enterprise-level data solutions for a wide range of applications (Datazenit 2014). The vast majority of websites that have a data requirement use MySQL to cater for that need.\textsuperscript{104} MySQL databases boast a number of technical advantages. First, they respond very quickly to client requests\textsuperscript{105} and this was an attractive offering given the reasons for the rejection of the first linkedarc.net server design outlined above. Secondly, they are eminently scalable. This means that you can build a MySQL solution, which initially deals with a relatively modest datastore size and simultaneous client number and then as the demand for the resource rises, there is no need to make any changes to accommodate the increased load.\textsuperscript{106}

\textsuperscript{104} A survey conducted in 2011 by bestvendor.com asked 500 developers about the tools that they use to create websites (Strom 2011). 56\% of those surveyed said that they used MySQL databases.

\textsuperscript{105} It is difficult to objectively quantify the speed of a typical MySQL database response. This lack of clarity is down to the number of variables involved both within the MySQL configuration itself and in the underlying server architecture (Oracle 2011). Facebook have started to migrate some of their servers over to a NoSQL configuration in the last number of years (Humble 2014). However, in 2010 when they were still using SQL as their primary datastore, it was reported by the company that they were able to service an individual SQL query in just 4msecs (High Scalability 2010). This equated to an ability to read a staggering 450M table rows per second.

\textsuperscript{106} MySQL scalability is not infinite. Extremely high-end requirements (for example, to write 40 million values per second) would not be achievable using a MySQL setup (Xaprb 2013). Having said that, it is unlikely that databases dealing with cultural heritage material would ever demand that level of service.
A third reason that MySQL databases are a popular choice is because of that very popularity. As a result of their being so widely used, there is a vast resource of online and print training material available and a well-established community of developers that are willing to answer questions about the workings of the system. Oracle, who now own the MySQL brand, also devote significant resources to its continued research and development and this leads to a healthy cycle of updates and bug fixes.

Unsurprisingly, due to its popularity, there is no difficulty in finding a Python library that enables a developer to interface with a MySQL database. linkedarc.net ver2 used the MySQLdb library. The key challenge that presented itself rested upon the fact that MySQL databases are relational in design, which is at variance with the RDF data model. Following the discussion of database types in Chapter 2, we know that there is a substantial structural difference between a relational database and an RDF triplestore. In order for linkedarc.net to provide RDF Linked Open Data resources, it was necessary to create a conceptual bridge (Figure 49) between the MySQL relational database on the one hand and the linkedarc.net client interfaces on the other, which were responding to and delivering RDF requests and responses.

The MySQL solution delivered a satisfactory performance when responding to queries that were received via the web service interface. However, and as will be explained in Chapters 6 and 7, it became increasingly apparent during the project that the linkedarc.net server needed to also provide a fully functional SPARQL interface. And were linkedarc.net ver2 to offer such an interface, it would have to be written from scratch. SPARQL is a relatively complex query language and writing such an add-on for linkedarc.net would have demanded a large investment of time and effort. On the basis

Figure 49: linkedarc.net ver2 bridging RDF and MySQL

107 MySQLdb (http://pypi.python.org/pypi/MySQL-python) and MySQL Connector/Python (http://dev.mysql.com/doc/connector-python/en/) are two options.
of this assessment, it was decided that the linkedarc.net datastore would need one final major revision.

**linkedarc.net version 3 – Apache Jena Fuseki**

The third and final version of the linkedarc.net server replaced the MySQL data backend with an Apache Jena RDF triplestore (Figure 50). Apache Jena allows you to create and query RDF graph data. The version of Jena used by linkedarc.net ver3 includes an add-on service called Fuseki, which provides a SPARQL interface to the Jena triplestore and in practice this is the only way that the linkedarc.net system updates and queries the data stored in the Jena datastore. Figure 51 shows how web service and URI requests come to be routed to the Apache Jena Fuseki endpoint via a SPARQL query constructor and these various scenarios will be described below.

---

108 At the time of writing, the latest version of Apache Jena Fuseki was ver2.3.0. See jena.apache.org for more information.
Serving SPARQL queries

The linkedarc.net SPARQL endpoint is available at http://linkedarc.net/sparql. Queries received by the Apache server and addressed to this URL are routed directly to the Fuseki engine, bypassing the linkedarc.net server code. The Fuseki engine allows the client to specify parameters and these are included in the URL that the client constructs. These parameters are summarised in Table 15.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>query</td>
<td>The SPARQL query.</td>
</tr>
<tr>
<td>output</td>
<td>This specifies the format that you want the data to be returned as. Valid values are json, text, xml, csv and tsv.</td>
</tr>
<tr>
<td>stylesheet</td>
<td>If ‘xml’ is the value specified for the output argument, you may choose to specify a stylesheet with which to style the XML data that is returned.</td>
</tr>
<tr>
<td>default-graph-uri</td>
<td>linkedarc.net data is stored within a single default graph in Apache Jena and as such this parameter should be left blank or omitted entirely.</td>
</tr>
</tbody>
</table>

Table 15: the linkedarc.net SPARQL endpoint parameter list

An example client request to the SPARQL endpoint is shown below.\(^{109}\)

http://linkedarc.net/sparql?query=SELECT+%2A+WHERE+%7B%0D%0A%09%3Fs+%3Fp+%3Fo%0D%0A%7D+LIMIT+10&output=json

This asks the Fuseki engine to parse and execute the SPARQL query shown in Table 16 and to return the results as JSON.

```
SELECT *
WHERE {
  ?s ?p ?o
}
LIMIT 10
```

Table 16: a SPARQL SELECT query to get the first 10 RDF triples in the triplestore

\(^{109}\) Note that URL parameters need to have certain invalid characters, such as spaces, replaced with valid characters. This is known as escaping the URL.
Serving URI requests

linkedarc.net resource

Additional notes on accessing the linkedarc.net URI resources

http://linkedarc.net/about/datauris

Consumers of the linkedarc.net system may wish to access specific data resources of the service. As the linkedarc.net server is Linked Open Data compliant, each of these resources is associated with a URI, which follows this format:

http://linkedarc.net/data/PROJECT/RECORD-NAME

Each of these URIs is also a URL as they are all dereferenceable web resources. This means that if a client sends a HTTP GET request to one of them, the linkedarc.net server will return a representation of that resource.

Figure 52 outlines how this happens in practice. A request for the resource, http://linkedarc.net/data/la_pp/context522, is received by the Apache server. This is routed to the linkedarc.net SPARQL query constructor, which constructs the query shown. This query is then sent to the Fuseki endpoint and the response is returned to the client. The format of this response depends on the form of the URL specified in the request. If the URL contains no extension, it will be returned in a JSON format.
Otherwise, and if the extension is recognised as being valid by the server, the server will return data in the corresponding format.\footnote{The following extensions are supported by the server: .json (JSON data), .xml (RDF/XML data), .sparqlxml (SPARQL/XML data), .ttl (RDF Turtle) .csv (Comma Separated Values data), .tsv (Tab Separated Values data) and .txt (text).}

**Serving web service requests**

**linkedarc.net resource**

Additional notes on the linkedarc.net web API

http://linkedarc.net/about/webapi

A web service can be thought of as a code resource that is located on a computer other than the application computer. Web services have become an incredibly popular way of exchanging information across networks such as the internet. They are typically used by application developers who wish to add certain network-based functionality to their services. For instance, a developer might want to include live weather data on their website. This can be achieved by making a call to an internet-based web service. A typical web service conversation goes as follows. The client computer sends a HTTP request to a server specifying the URL of the web service that it wants to interact with and the list of parameters relevant to that request. The server then responds with a packet of information, which tends to be formatted as JSON.

Figure 53: the processing of a web service call by linkedarc.net v3.0.0

Figure 53 describes how the linkedarc.net server deals with incoming web service requests. A request is received and it is interpreted by Apache’s routing module to be destined for a linkedarc.net web service method. Apache duly passes the request on to the web.py Python module, which in turn directs it to the linkedarc.net server module that handles web service requests. If valid, this module passes the request on to the
function that deals with the particular request type. This calls the SPARQL query constructor module, which creates a SPARQL query based on the request. This query is then sent on to the Fuseki endpoint and once it receives a response from Fuseki, it returns the data to the client in JSON format.

linkedarc.net ver3.0.0 exposes a number of web service methods and these are listed alongside their parameters and a short description in Table 17.

<table>
<thead>
<tr>
<th>Web service method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>ws/query?request=projectlist</code></td>
<td>Get a list of all the linkedarc.net projects.</td>
</tr>
<tr>
<td><code>ws/query?request=projectontologyurl&amp;project=[PROJECT_ID]</code></td>
<td>Get the URL of the ontology file for the linkedarc.net project.</td>
</tr>
<tr>
<td><code>ws/query?request=datatablelist&amp;project=[PROJECT_ID]</code></td>
<td>Get a list of all the datatables in the linkedarc.net project.</td>
</tr>
<tr>
<td><code>ws/query?request=recordlisthtml&amp;project=[PROJECT_ID]&amp;datatable=[DATATABLE_ID]&amp;range_from=[RECORD_INDEX_START]&amp;range_to=[RECORD_INDEX_END]</code></td>
<td>Get HTML that lists certain values in a range of datatable records. Note that linkedarc.net creates this HTML document using a schema for the requested datatable that contains predicate value placeholders.</td>
</tr>
<tr>
<td><code>ws/query?request=recordhtml&amp;project=[PROJECT_ID]&amp;datatable=[DATATABLE_ID]&amp;record_id=[RECORD_ID]</code></td>
<td>Get the HTML that is filled in with values linked to the requested datatable record. Note that linkedarc.net creates this HTML document using a schema for the requested datatable that contains predicate value placeholders.</td>
</tr>
</tbody>
</table>

111 Note that these web service method URLs have been abbreviated for legibility. Prefix each with `http://linkedarc.net/` to create its complete form. Text entries contained within square brackets are placeholders for web service method argument values. For example, `http://linkedarc.net/ws/query?request=projectontologyurl&project=la_pp`.

112 Display schemata are discussed in the context of the linkedarc.net web app later in this chapter.
ws/query?project= request=search
&project=[PROJECT_ID]
&datatable=[DATATABLE_ID]
&record_index=[RECORD_INDEX]
&searchby_param1=[SEARCH_PARAM1]
&searchby_comparison1=[SEARCH_COMPARISON1]
&searchby_value1=[SEARCH_VALUE1]
&searchby_valuetype1=[SEARCH_VALUETYPE1]
&searchby_constraint1=[SEARCH_CONSTRAINT1]

Search the datatable on the basis of a set of search criteria.

<table>
<thead>
<tr>
<th>Table 17: linkedarc.net v3.0.0 web service methods</th>
</tr>
</thead>
</table>

**Client types**

Four clients are listed in the network diagram shown in Figure 50. These are examples of the types of client that the linkedarc.net system is capable of servicing as a host. linkedarc.net was designed to be enormously flexible as a data provider. This means that many different types of client can consume its data. This approach follows one of the basic tenets of the Linked Open Data philosophy – to focus on the data itself and not on the applications that will use that data (ablvienna 3013). The ways that a dataset eventually comes to be used are determined by the data consumer. The objective here is to encourage imaginative and novel uses and re-uses of data, leading to unforeseen and valued outcomes.

![web app](image)

**Figure 54: a web app linkedarc.net client**

The first client is the linkedarc.net web app (Figure 54) and will be discussed in detail later in this chapter. The second client is another web app and it is included as an example of this open-endedness that is referred to above. This second client might be developed by a third party, which is interested in consuming only a subset of the overall service offered by linkedarc.net. The pptextiles.linkedarc.net project is one example of such an approach and we will return to this project in Chapter 5.
The next client type (Figure 55) is an iOS application, which communicates with the linkedarc.net system via its web service or SPARQL interfaces. The final client (Figure 56) is labelled a ‘data mining exercise’ and represents a once-off server querying session. For example, let us imagine that we wish to find out the number of people who worked at the site of Priniatikos Pyrgos in 2010. We might then compare this set of data to other data that has been mined from other archaeological site datasets. Data mining exercises now represent an important tool available to the digital humanist (Kirschenbaum 2009). A typical data-mining scenario begins with a research question. Helper applications such as Open Refine\textsuperscript{113} are then used to retrieve data from a source like linkedarc.net and finally the mined data is cleaned up and passed on to a data analysis tool such as Excel. The outcomes of these analyses go towards the answering of the original research questions and more often than not they also go on to fuel the asking of new questions, thereby repeating the process.\textsuperscript{114}

\textsuperscript{113} Open Refine is discussed in more detail in Chapter 5.

\textsuperscript{114} See Chapter 6 for examples of this research method type in practice.
The linkedarc.net web app

If you launch a browser and enter linkedarc.net into its address bar, the linkedarc.net web app is what you will see. In itself, the web app does not deal with the management of any RDF data. The two sets of logic that manage the server and the web app are so entirely separate, that if required, they could be easily run on two separate servers.\textsuperscript{115} The web app is a consumer of linkedarc.net’s services and as we have seen it is just one of many clients that can consume linkedarc.net services. The amount of access that the linkedarc.net web app has to the data stored in the server’s Apache Jena triplestore is determined by the server. In this sense, the server acts as the data’s gatekeeper.

As a basic objective, the linkedarc.net web app needs to give a representation of the linkedarc.net server’s data. As a second requirement, it needs to allow the user to find data within the dataset that is of interest. Third, the linked nature of the linkedarc.net server data needs to be emphasised and displayed appropriately to the user. Fourth, users should be able to conduct statistical analyses of the data. The results of these analyses should be represented informatively. Finally, users should be able to access the web app services on a range of web-enabled devices.

\textbf{What is a web app?}

A web app is defined as ‘an application in which all or some parts of the software are downloaded from the Web each time it is run. It may refer to browser-based apps that run within the user's Web browser, or to "rich client" desktop apps that do not use a browser, or to mobile apps that access the Web for additional information’ (PCMag 2015). Since the advent of Web 2.0, web apps have become an increasingly popular way of providing computer-based functionality to a user. In fact, it could be argued that the need to give it its own sub heading within the wider application pool may be now anachronistic, as applications that do not follow its definition are becoming increasingly uncommon.

\textsuperscript{115} In practice, for logistical reasons (cost and server administration), they run on the same server, which is hosted by Digital Ocean (http://digitalocean.com)
The web app development domain is currently divided between examples that do the majority of their work on the server and those that more extensively employ the processing power of the client (Figure 57). The linkedarc.net web app falls into this latter category. While PHP or ASP.NET web apps will prepare pages on the server based on a set of parameters that have been provided by the client, client-side web apps like linkedarc.net use Asynchronous JavaScript and XML (AJAX) functionality, which is embedded within the JavaScript and jQuery framework, to divide the workload more equally between the client and server.\footnote{http://jquery.com}

While most people are familiar with the idea that a web browser can download documents and other web resources from a server using HTTP, it is less commonly known that a web app may also employ HTTP to request smaller amounts of data as and when they are needed by the user and not only when a web resource is first loaded. This type of client-server interaction uses network bandwidth more efficiently as pages do not need to be constantly refreshed whenever the user requests an update. Instead, a combination of AJAX and a dynamic user interface update library such as jQuery, allows the developer to leave the majority of the web page unchanged, while a small amount of data is requested, downloaded and then used to update the page content. The content

\footnote{In fact, the linkedarc.net web app does use some PHP technology but this is very limited. It is solely used to allow for the passing of state information between pages.}
presented to the user on a linkedarc.net web app page is either downloaded when the page is first loaded or it is dynamically retrieved in real-time using AJAX technology. This approach was pioneered by the Google Gmail app (Zucker 2007).

**Consuming linkedarc.net web services**

![Diagram of AJAX web service call]

Figure 58: linkedarc.net making an AJAX web service call

In practice the majority of RDF data manipulated by the web app is accessed using the linkedarc.net server’s web services. Figure 58 describes how a typical web service call is initiated as a result of a user intervention and how the response to that web service call is dealt with by the web app. For example, a user might enter some text into a field on a page and then click on a button labelled ‘search’. The web app constructs an AJAX call based on the user’s request, sends that request to the linkedarc.net server’s search web service. The server responds with a JSON packet, which the web app parses and then proceeds to display this information as an update to the content of the page.

**The basic design**

The linkedarc.net web app user interface design follows a basic set of design principles that are used by many data-driven sites on the World Wide Web. The layout of data-heavy sites such as Wikipedia and indeed all of the Linked Open Data compliant sites reviewed at the beginning of this chapter, exhibit design practices, which have been honed over two decades of website development (Ruluks 2014). Users of web-based
content have become familiar with engaging with web data in a particular way, and, therefore, it makes sense for the linkedarc.net web app to capitalise on this familiarity of practice.

![Figure 59: the linkedarc.net web app's responsive layout](image)

The layout of every linkedarc.net web app page is responsive.118 Layout types determine how a web site reacts to changes in the size of its container, which is typically a browser window. Responsive web sites combine elements of a liquid and an adaptive design by resizing page elements relative to the width and height of the screen but also by adjusting whole sections of a page whenever certain width and height thresholds have been passed. Figure 59 shows how the layout of the project page reacts to changes in the browser’s window size. On the left, the page displays more vertical content to the user. This changes in the mobile layout, as vertical reach is sacrificed in order to maintain text size and, therefore, legibility.

The page header contains links to the home page and to each of the other pages that make up the site. These links are accessible via a dropdown menu (Figure 60).

---

118 Typically, there are acknowledged to be four basic web layout design types: fixed or static, liquid or fluid, adaptive and responsive (Pettit 2014).
Chapter 4 - linkedarc.net

The linkedarc.net web app employs a ‘stateless’ operation, in which each new page’s state is reinitialised as a function of its page parameters (Zuzak et al. 2011). This means that a linkedarc.net web page URL can be used as a permanent link to its content. This also means that the browser’s history function can be used to navigate between pages.

One final point to make about the linkedarc.net web app is that for each project that it hosts, there is a corresponding project administrator. The administrators have control over the visual representation of the data that is sourced from the linkedarc.net server. Because of this flexibility, a pair of linkedarc.net projects might be represented in two entirely different ways by the system.

Home page

linkedarc.net resource

The linkedarc.net web app home page

http://linkedarc.net

The first page presented to the user when they enter the web app is the home page (Figure 61). A popup display lists the latest news items and these include announcements about new features, new videos, new help guides, and so on. Clicking anywhere on the page dismisses the news popup and the user is presented with a world map showing the locations of the linkedarc.net projects. Clicking on any of these markers displays another popup box, which gives a brief introduction to the project in question (Figure 62). Clicking on this popup navigates the user to the corresponding project page.
Figure 61: the linkedarc.net web app home page

Figure 62: navigating the linkedarc.net web app’s home page map

Figure 63: the linkedarc.net web app project page
Chapter 4 - linkedarc.net

Project page

linkedarc.net resource

The linkedarc.net web app Priniatikos Pyrgos project page

http://linkedarc.net/projects/project.php?project=la_pp

Figure 63 reproduces the project page. Projects, as has been explained in our overview of the linkedarc.net server (Figure 47), are divided into datatables, which in turn contain the project’s record entries. This page lists the datatables linked to the current project. This particular project includes datatables called ‘Catalogued Objects’, ‘Contexts’ and ‘Images’, and so on. The page includes a small description of each datatable and a number indicating the number of records contained within.

Lastly, the project page contains a link to the ontology file for the respective project data. The project’s administrator decides upon the form of this serialisation but it is likely that most linkedarc.net project ontologies will be rendered using the OWL format, which has been outlined in Chapter 3. A central tenet of the linkedarc.net philosophy is the promotion of the re-use of its Linked Open Data assets. In order to achieve this, the baseline requirement is that the project dataset be made as transparent to interested parties as is possible. Linked Open Data transparency begins with the opening up of a dataset’s internal structure, and this is the reason that the project ontology file is displayed so prominently within the linkedarc.net web app design. OWL serialisations can, however, be complicated documents. In order to make sense of an OWL serialisation, an interested third party must be reasonably well versed in the technical intricacies of RDF and the OWL model, and this level of competence demands a prior and not insubstantial investment on their part. As such, it needs to be acknowledged that the publication of a dataset’s ontology does not in itself render that dataset re-usable by

119 The amount of datatables in a project and the form of the records contained within those datatables is determined by the project administrator, who is referred to throughout the following text in the context of how data elements are displayed to the user. Project administrators play an important role in the management of project data on the linkedarc.net system. They are the point of contact between the linkedarc.net team and the project data creators and so following Kansa and Kansa’s (2013) data publication model, they play the role of the data author, highly familiar with the data’s content and structure but potentially less au fait with the intricacies of Linked Open Data practice.
an unlimited community of users. We will return to this topic of Archaeological Semantic Web user requirements in Chapter 7.

**Datatable page and display schemata**

**linkedarc.net resource**

The linkedarc.net web app Priniatikos Pyrgos Catalogued Objects datatable page

http://linkedarc.net/projects/project_datatable_browse.php?project=la_pp&datatable=catobj

![Figure 64: the linkedarc.net web app datatable page](image)

The datatable page displays a list of the records contained within the selected datatable (Figure 64). The form of this display is determined by a display schema file, which has been created by the project’s administrator for each datatable.

```html
<!-- LINKEDARC.NET LIST DISPLAY SCHEMA FILE

PARSING STEPS:
1) REPLACE ANY PREFIXES IN TEXT WITH ITS VALUE
2) REPLACE ANY TEXT STARTING WITH [ AND ENDING WITH ] WITH THE PREDICATE (FOR :predicate)
OR THE CORRESPONDING PREDICATE'S OBJECT (:object)
-->

@prefix la: <http://linkedarc.net/ontology/>.
@prefix la_pp_ont: <http://linkedarc.net/ontology/la_pp/>.
@prefix la_pp: <http://linkedarc.net/data/la_pp/>.
@prefix dct: <http://purl.org/dc/terms/>
```
An example linkedarc.net display schema file is reproduced in Table 18. The format used for these files is a subset of the HTML format. It is, however, proprietary to the linkedarc.net system. It essentially combines the elements of the HTML format with some new elements (prefixes, repeat blocks and placeholders), which handle the insertion of data into the schema and these are similar to those found in the RDF Turtle format. The processing of a display schema file can be summarised as follows:

1. The web app makes a web service request to the server for the HTML content of a project’s datatable. As part of the arguments for this web service call, it gives the server a list of the records, whose data is to be included in the returned HTML.

2. The server finds the relevant display schema and expands any prefixes found within the placeholders. A placeholder is a section of text demarcated by square brackets.

3. The [REPEAT-START] and [REPEAT-END] markers indicate that the block of HTML code that they contain should be duplicated for the number of records requested.
4. The placeholders within the [REPEAT-START] and [REPEAT-END] blocks are then replaced with the data associated with the list of requested records. Each placeholder contains an RDF predicate/object instruction, which is resolved for each individual record. For example, [rdfs:name:object] is replaced with the object value that is associated with a record via the rdfs:name predicate. These placeholders are capable of encoding complex predicate/object chains as is shown in Figure 65.\textsuperscript{120}

5. Once all of the placeholders have been replaced with data, the HTML code is returned to the web app, which displays it to the user.

\textsuperscript{120} Note that the placeholder included in the Figure 65 example, ecrm:P89_falls_within(la_pp_ont:LA_E3_Trench):rdfs:name:object], contains a URI in curved brackets after the predicate, which means that the parsing engine should only look at object values that are of the type specified by the URL in the curved brackets.

Lastly, the web app also handles the pagination of the datatable’s records allowing users to browse through pages of record lists. Users can navigate to the record page of any record by clicking on any one of its values.
**Record page**

**linkedarc.net resource**

The linkedarc.net web app Priniatikos Pyrgos Catalogued Object 05-1006 record page

http://linkedarc.net/projects/project_datatable_record.php?project=la_pp&datatable=catobj&record_id=05-1006

The record page is the centre of the linkedarc.net web app, as it is here that the majority of the linkedarc.net server’s data is presented to the user. Record pages also use display schema files to deal with the layout of a particular record’s data. The rules governing the creation of record layouts is very similar to that which has already been described in the context of the datatable display schemata, with the exception that a record display schema is only concerned with the display of a single record’s data at any one time. Record display schema files also allow for the inclusion of SPARQL blocks. These are contained within `[SPARQL-START]` and `[SPARQL-END]` markers and allow the user to construct complex queries with which to retrieve the data related to a record. SPARQL blocks can be useful when wanting to add values that summarise other values and we will see this in action when we discuss the linkedarc.net display of Priniatikos Pyrgos content in Chapter 5.

Figure 66 shows how the record’s data has been divided up into a series of tabs (General, Analysis, et cetera). This is a convenient way of categorising information for the user. The form that each datatable’s record takes, however, is entirely determined by the judgement of the project’s administrator and is not decided upon by the overall system design (Figure 67). A project administrator might, for example, decide to represent a datatable’s records as graph or map data and not as a table. As long as the particular layout conforms to the HTML standard, then it can be included in the linkedarc.net web app as a record page.
Figure 66: the linkedarc.net record page

<table>
<thead>
<tr>
<th>General</th>
<th>Analysis</th>
<th>Object specific</th>
<th>Images</th>
<th>Workbench</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catalogue code</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specialist</td>
<td>mnse</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Period (preliminary analysis)</td>
<td>Classical Period</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Date excavated</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Links</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trench</td>
<td>g1000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contexts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sub-contexts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Locs</td>
<td>g1001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pails</td>
<td>g1001-02</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 67: placement of the display schema HTML in a linkedarc.net web app record page
Links are a key aspect of the linkedarc.net web app model. They put into practice and reflect the underlying links inherent in the source RDF data. But again it needs to be emphasised that the amount and character of these links is determined by the project administrator, who creates the display schema files. In the example shown in Figure 66, the predicate names listed on the left-hand side each link to their corresponding predicate URI. For object values that are in the form of URIs, it is also common practice that these also contain links to their values. Links encourage a user to explore the full extent of a Linked Open Data network of information, and it is through this process of discovery that any Linked Open Data dataset comes to be understood by the user.

Figure 68: RDF graph data with multiple objects linked via the same predicate

Figure 69: processing a linkedarc.net display schema file
Object values may come in the form of a single value or they may evaluate to a set of values. Consider the diagram shown in Figure 68. The record subject at the centre of the diagram is linked to a single object value via the rdfs:name predicate but it is linked to several objects via the same rdf:type predicate. This, as we have seen, is an entirely valid implementation of RDF.

linkedarc.net supports this scenario by allowing project administrators to add [REPEAT-START] and [REPEAT-END] blocks to their display schemata. If encountered, the display schema file parser duplicates the HTML code contained within these blocks for each value in the set that is found to be associated with the subject via the specified predicate (Figure 69).

Object value types are also very important and need to be taken into account when managing RDF data as different types of data require the use of different user interface elements, particularly when it comes to searching and updating data. For instance, strings, integers, decimals and dates are all formatted differently and this needs to be reflected in the user interface design. In that context, the display schema files provide a mechanism to allow the project administrator to communicate this data type information to the web app with the inclusion of a ‘data-rdf-objecttype’ HTML attribute (Smith 2013). The web app can then use this information when it needs to construct search and update interfaces for a record.

![Tags](http://linkedarc.net/vocabs/loomweight–pyramidal)
![Tags](http://linkedarc.net/vocabs/material–ceramic)
![Tags](http://linkedarc.net/vocabs/object–weight)
![Tags](http://linkedarc.net/vocabs/object–loomweight)

**Figure 70: SKOS concepts as object values in the linkedarc.net web app**

SKOS concepts are a particular and extremely important type of object value, which were introduced in Chapter 3. They are supported by the linkedarc.net web app with the inclusion of the ‘tag-set’ CSS class (Bos et al. 2014). Object values, which are contained within an HTML element with the ‘tag-set’ class are displayed to the user as is shown in Figure 70.
One of the major advantages of using SKOS concepts is that they allow for the identification of patterns across datasets and the linkedarc.net web app provides a mechanism to do this using the Workbench user interface element. Project administrators can choose to include a Workbench tab in a record display schema file. The SKOS values reproduced in Figure 70 include a globe icon to their right and when clicked, this will display the ‘Start exploring the Semantic Web’ popup (Figure 71).
This popup allows the user to search for matching Linked Open Data resources on other cultural heritage Linked Open Data services and to add these matches to the Workbench tab (Figure 72). It is worth stressing the importance of the linkedarc.net Workbench feature in the context of this thesis. It essentially realises the central tenet of Linked Data: that any one node of information can be related to any other information node, regardless of where they both reside – i.e. within the same dataset or not. Currently, linkedarc.net supports the querying of Linked Open Data resources on the Europeana and British Museum datastores. The linkedarc.net Workbench can be thought of as a form of digital pencil and paper. It can be used to record the researcher’s observations as they go about the business of mining the linkedarc.net dataset. It allows the user to note down the patterns that emerge within the linkedarc.net data and the relationships between those data and data residing in the wider Semantic Web.

A key requirement of any Archaeological Information System is an ability to search within its dataset. The linkedarc.net web app provides search capabilities in a number of

---

121 Note that the content added to the Workbench is temporary and if the page is reloaded, it will be removed.
122 Allowing a digital researcher to take notes while they navigate through the system’s dataset is becoming an increasingly common feature of what are now known as Virtual Research Environments or VREs (De Roure et al. 2009, Carusi & Reimer 2010). While linkedarc.net lacks the core functionality (most notably, an ability to support collaborative research) that would allow it to claim membership of this category of applications, the workbench feature is an exploratory gesture in the direction of this more interactive form of digital research practice.
different ways but one of the most powerful is the search feature built into the record page. Clicking on the search icon at the top of any record page puts the page into search mode (Figure 73). This has the effect of replacing all object value fields on the page with input fields, which the user can then enter search criteria into.

Object value types are important in this context, as certain object types will enforce particular formatting rules on the text that the user enters. For example, if the user chooses to search within a date field, they will need to select a date using a popup calendar. Options appear alongside each text or number field, which allow the user to specify how they would like the value to be compared. For example, does the user want to consider only exact text matches? Does the user want to retrieve all of the records that have a number value in a field that are less than, equal to or greater than a particular search value?

The user clicks the ‘Start find’ button to begin the search. A query is constructed based on all of the criteria that have been entered. These criteria are combined together using the logical AND operator. The linkedarc.net server returns a results set that matches the search criteria. The user can then step through this list using the record navigation control located in the top-right-hand corner of the page.

By default, each record page also lists a number of alternative representations for the data in question. Users have access to a SPARQL/XML, RDF/XML (with and without XML style-sheet), RDF-Turtle, JSON, plain text, CSV and TSV serialisation of any record. Users can also view the record using either the LodLive viewer or the Q&D RDF browser.123

---

123 The LodLive tool is discussed in Chapter 8.
Faceted search page

linkedarc.net resource

The linkedarc.net web app faceted search page

http://linkedarc.net/projects/search

Besides the search interface contained within the record page, linkedarc.net also provides a dedicated search page and this is shown in Figure 74. The range of options available to the user changes as they select search options, which narrows down their overall search set. For example, selecting a project from the project list restricts the data range to the contents of that project. A user can also choose to filter by datatable and by the predicates used by records within that datatable.

Figure 74: the linkedarc.net faceted search interface

Figure 75: adding criteria to the linkedarc.net search interface
When searching across all project data, within an entire project or a datatable within a project, the user simply has to enter a single search term. If, on the other hand, the user wishes to search at the level of a record’s predicate, they can choose to include multiple search criteria as is shown in Figure 75.

![Figure 76: linkedarc.net search results](image)

The results of a search operation are listed below the search criteria fields in a table of triples (Figure 76). These may be paginated depending on their total number. Users can narrow down the number of results by choosing to search again within the results set. The user can click on any resource’s subject and this will bring them to the corresponding record page.
**SPARQL query editor**

**linkedarc.net resource**

The linkedarc.net web app SPARQL query editor

[http://linkedarc.net/sparql](http://linkedarc.net/sparql)

The SPARQL query editor\(^{124}\) provides advanced querying of the linkedarc.net dataset (Figure 77). The syntax of SPARQL has already been introduced in Chapter 3 and we will consider at greater length the rationale for choosing to use SPARQL as an interface into an archaeological dataset in Chapters 6 and 7. In this coming section, we explain how SPARQL can be employed to access content stored within the linkedarc.net data ecosystem.

![Figure 77: the linkedarc.net SPARQL query editor](image)

Various options are offered to the SPARQL query writer and these are shown in Figure 78. Clicking on a prefix button will add the corresponding prefix line as formatted SPARQL text to the editor box. The ‘Sample queries’ provide a useful introduction to both the SPARQL language itself and to the structure of the linkedarc.net data. Selecting a query description from this list inserts the corresponding SPARQL text into the editor box. The ‘Output’ field lists all of the data serialisations supported by the server and if

\(^{124}\) The SPARQL query editor uses the YASQE plugin ([http://yasqe.yasgui.org/](http://yasqe.yasgui.org/)) to provide advanced user interface features, which have been specifically designed for the writing of SPARQL queries within a web environment.
you select ‘XML’ from this list, you can also select a style sheet with which to style the resultant RDF/XML response. Clicking the ‘Get Results’ button sends the query to the server and once processed, the results will be displayed in the format specified.\footnote{Note that if an error is contained within the query, the server will provide some suggestions as to its possible cause.}

Conclusions

For readers with some experience of web technologies or of the Semantic Web more specifically, the preceding account will have come across as remarkably familiar. All of the approaches described (with the exception of the linkedarc.net web app’s use of proprietary display schemata) are well-established methodologies of these fields. In that case, one might reasonably be drawn to ask, why it was necessary to build the linkedarc.net system in the first place and why it was important to describe that process in this chapter? The answer to the first question is that the linkedarc.net server and web app represent crucially important tools of this research project. We will see in the next two chapters how linkedarc.net comes to be used as the host for the RDF data of the Priniatikos Pyrgos Project and as a medium through which this data is accessed by the archaeological consumer. To put it simply: this written work could not have existed without linkedarc.net. It is the project’s test bench and its medium to ‘methodological understanding’ (McCarty 2005 p. 120). In answer to the second question, it was necessary to describe the process of building linkedarc.net because it is important for you, the reader, to understand the production side of RDF, if you are to fully appreciate the entire breadth of the Archaeological Semantic Web.
Chapter 5 – Populating linkedarc.net

Creating an RDF archaeological dataset

‘What men really want is not knowledge but certainty.’ (Bertrand Russell)\textsuperscript{126}

Introduction

In the last chapter, I introduced the linkedarc.net server and web app, as two components of the one Linked Open Data Archaeological Information System. This account was intentionally focussed on the technical aspects of the system’s design and for the most part it avoided any discussion of real archaeological data. This is the task that I turn to now, as I consider the evolution of the Priniatikos Pyrgos dataset from private FileMaker project to public RDF resource housed on the linkedarc.net server. RDF data mapping and publishing can be a laborious and time-consuming process, particularly when it involves a complex mesh of information as is evidenced in the Priniatikos Pyrgos example. It is a hurdle that many Linked Open Data archaeological projects struggle to overcome in the sense that it is multi-staged, tedious and it demands enormous levels of human oversight. As such, one of the primary functions of this chapter is to illustrate in no uncertain terms just how challenging RDF archaeological data creation can be. Also, while the many processes involved are undoubtedly technical, I will show how no data operation is ever entirely removed from a theoretical oversight (Shanks et al. 1987 p. 9).\textsuperscript{127}

\textsuperscript{126} This aphorism is generally attributed to Russell but it was never included in any of his official publication.

\textsuperscript{127} Note that this chapter has been structured so as to abstract the sequence of actions that actually led to the creation of the Priniatikos Pyrgos RDF dataset. This has been done to make the narrative more comprehensible for the reader. The reality of the mapping process involved substantially more to and fro between the building of the system described in Chapter 4 and the creation of the RDF content to populate that system.
Priniatikos Pyrgos is a small site (c1.26 ha area) located on the southern shore of the Mirabello Bay in East Crete (Figure 79). It is situated on a limestone coastal promontory just a short walk from the modern town of Istron/Kalo Chorio. In this part of the island the coast is separated from the mountains by a thin strip of plain, which over the millennia has served as an important east-west communications corridor along Crete’s northern shoreline. The site is just a few miles to the west of an isthmus, which connects the north coast to the south coast linking the towns of Pachia Ammos and Ierapetra. The area surrounding Priniatikos Pyrgos has been densely populated throughout history with

---

128 Arriving at an estimate of Priniatikos Pyrgos’s size is problematic for a number of reasons. Firstly, it is highly likely that large areas of the site remain uninvestigated. It has been speculated on the basis of geophysical survey work carried out within the confines of a nearby football pitch, that during the Bronze Age, Priniatikos Pyrgos was a small part of a much larger urban complex, which is now largely covered by the thick alluvial deposits laid down by millennia of action by the Istron River. Also Priniatikos Pyrgos is not simply a single period site. It contains evidence of occupation from a number of different periods, each with its own size characteristics. In other words, a medium sized Bronze Age site does not necessarily equate to a medium-sized Iron Age site and so on. Having said that, based on Whitelaw’s (2001 fig. 2.10) analysis of a range of Cretan Bronze Age sites and Hayden et al.’s survey of the general district of the western Mirabello Bay region (2004 p. 46), Priniatikos Pyrgos can be comfortably read as being small-scale for most periods. The 1.26ha area figure represents the area of excavated space at the site and does not take into account the areas that remain uninvestigated on its extremities.
small mostly rural settlement and it seems likely that Prinatiikos Pyrgos along with the adjacent headland of Nisi Pandeleimon served as an urban nucleus for this local settlement hierarchy (Hayden 2014).

The site first came to prominence in an archaeological sense in the early 20th century when the pioneering American archaeologist, Edith Hall, sunk a test trench somewhere within its environs (Betancourt 2014 p. 11). After this brief yet promising investigation, it lay undisturbed for more than a century, when in 2005 Tsipopoulou and Hayden began a rescue excavation, which followed the Vrokastro Area Survey Project (Hayden 2014 p. 15). This rescue project was backed by the American School at Athens and excavation continued for a further season until 2006. The following year, the site’s management was handed over to the care of the Irish Institute of Hellenic Studies at Athens and fieldwork continued until 2010.129

Remains of human occupation spanning over 5,000 years were uncovered during the 2005-2010 excavation seasons. The first major period of occupation stretched from the Final Neolithic (c.3500-3100 BCE) until the end of the Bronze Age (c.1100 BCE) (Hayden & Tsipopoulou 2012; Molloy et al. 2014 p. 1). The site was re-inhabited during the Late Geometric, Classical, Hellenistic and Early Roman periods (Hayden 2014 pp. 20–21). A short hiatus drew to a close in the 6th century, when ecclesiastical buildings came to dominate the site and the surrounding area until their collapse and abandonment in the 9th century (Bridgeford et al. 2015). More recently, there are suggestions that the headland may have housed a small chapel and the recent excavations uncovered evidence that at least part of the site was used as a gun emplacement during the Battle of Crete in World War II.

The archaeological recording systems used at Prinatiikos Pyrgos

The Prinatiikos Pyrgos rescue project recorded its activities using the Locus-Pail (LP) archaeological system in keeping with the approach taken by the majority of North American-influenced projects operating in the Eastern Mediterranean region (Morgan 2010).130 In this system, locus numbers are assigned to spatial areas that are delineated by architectural boundaries, most commonly walls, or perhaps absences of cultural material.

129 Since the cessation of excavation, the project has been engaged in a series of study seasons, which continue to the present year (2015).
130 The LP system has its origins in the ‘Wheeler box’ method and can also be referred to as the locus-basket system (May et al. 2015 p. 3).
Pails subdivide loci into administrative and, to a lesser extent, stratigraphic units (Pavel 2012 pp. 49–50). For example, let us imagine that we define a Locus 100 as the material excavated from the interior of a walled space as is shown in Figure 80. The pails 100-001, 100-002 and 100-103 represent each stratigraphic unit within the area of Locus 100.

![Figure 80: a cross-section describing the Locus-Pail system of excavation and recording](image)

This understanding of the pail is, however, not its only interpretation. In practice, pails tend to be used to allow for the excavation of an archaeological area by inexperienced archaeologists. In this case, a pail equates to a ‘spit’, which is an arbitrary block of soil (usually c.7cm in depth) excavated from a locus. Spit excavation places very few interpretive demands on the archaeologist and it runs the risk of missing subtle small-scale stratigraphic change.

The change of management at the site in 2007 also saw the introduction of a new excavation and recording system and this was employed within all new areas of investigation. Single context (SC) recording is the standard methodology used by commercial archaeological units in the UK and Ireland, as well as to a lesser extent in parts of mainland Europe and in pockets across the rest of the world (Morgan 2010). The core conceptual unit in the SC method is the context. ‘Any single action, whether it leaves a positive or a negative record within the sequence, is known as a “context”’ (Museum of London 1994 sec. 1.2). SC is intrinsically tied to site planning and stratigraphic analysis, so much so in fact, that in the Museum of London’s SC guide, which is acknowledged as the authority on SC, it is referred to as the ‘Single Context Planning System’. In this respect, SC’s ultimate goal is to represent a site’s excavated material on a plan drawing and as a network of stratigraphic relationships known as a Harris Matrix (Harris 1975).

---

131 Certain trenches, such as Area H, started before 2007 and, as a result, recorded using the LP system, continued to use this system in the post-2007 era.

132 A Harris Matrix is a schematic representation of the chronological relationships between contexts. Contexts on a Harris Matrix are linked using lines. If a context is linked to the top of
The Priniatikos Pyrgos paper record

As a result of the site’s varied administrative history, the data produced during the excavation of Priniatikos Pyrgos is a hybrid mixture of the LP and SC systems. To a large extent, whether an archaeologist chooses to excavate and record using one system and not another does not greatly affect the form that the project’s data takes. The format of the recording of a context and a locus is typically quite similar. Therefore, while both these conceptual entities should not be confused as semantically equal, at a data level, they are certainly compatible. We will now consider the range of data entities used at the site of Priniatikos Pyrgos in detail, looking at the rationale for each entity’s inclusion in the system and its association with a particular category of archaeological investigation.

The first point to make is that all of the Priniatikos Pyrgos records, excepting the photographs, which were captured digitally, were created using paper, pen and pencil. Figure 81 summaries the various categories of paper record and their conceptual links to another context, the implication is that this second context predates the first. If the first context links to the side of the second, they are both considered to be contemporaneous.
one another; essentially all of the paper records are interlinked using record identifiers. While the majority of these record types will be familiar to most archaeologists, some are unique to the site (the sub-context being the most obvious example of this) and others are unique in their detail. This pattern of conceptual similarity and divergent implementation has been highlighted in the introduction and by other commentators (May et al. 2015 p. 3) and represents one of the principal challenges when digitising archaeological record sets.

Figure 82: the Priniatikos Pyrgos C676 context sheet

The context and the locus record sheets form the nucleus of the Priniatikos Pyrgos model. Context sheets come in a number of different forms, each being used to record a different type of context. The most commonly used at the site was the ‘Fill/deposit/spread/soil’ context sheet, which is reproduced in Figure 82 for the context C676, and there were also other more specialised sheets used to record skeleton, masonry and cut contexts. The C676 sheet is a good example of a typical Priniatikos Pyrgos context sheet. It represents a deposit abutting one of the many limestone bedrock

---

133 A range of schemata was used to format the identifiers of each record type. For instance, a context 101, would be labelled C101, while a sub-context contained within C101 would be labeled 002-0323, where 2 refers to the trench number. While each record schema tends to be different in form, they all share the basic requirement that all identifiers be unique.
outcroppings encountered at the site. The front of the sheet is divided into broad data categories: administrative information (name of context, excavators involved, photo and plan numbers, et cetera), physical measurements, soil observations, stratigraphic relationships, finds and samples and a sizeable discussion field. The reverse is devoted to the planning of the context. A smaller sketch plan situates the context spatially relative to other contexts and to the parent trench. This is accompanied by a schematic Harris Matrix section, in which the recorder can relate the context chronologically to other contexts. The remainder of the back page contains a large grid, within which a detailed SC plan and/or section drawing can be drawn (Museum of London 1994).

One major particularity of the Priniatikos Pyrgos system is its use of a record type called the sub-context. Sub-contexts were employed by the project to subdivide the recording of a context excavated over a number of days into separate records. The thinking behind this was firstly administrative. It allowed for larger contexts, such as topsoil, to be broken into smaller more manageable units. Secondly, the sub-context was used as a safety mechanism protecting against failures to recognise new stratigraphic layers within the soil matrix, and given that the site was excavated predominantly using the labour of inexperienced student archaeologists, the use of this type of precaution was not unwarranted.

Sub-context records placed more of an emphasis on the material excavated from the context than their context sheet equivalents (Figure 83). They contain similar administrative information (sub-context name, parent context, trench, excavators, et cetera) to that of the context sheet. In respect to finds, the sheet contain fields, which allow the archaeologists to list the number of bags of ceramic, human bone, animal bone, shell, carbon and any other category of artefact or ecofact that were recovered during the sub-context’s excavation. Sub-context sheets were also used by the project apothēkē manager as an auditing device to ensure that all material excavated during any one day at

---

134 SC plans were first seen in the ‘open-air’ excavations conducted by Philip Barker, particularly at the site of Wroxeter in England (Barker 1982 p. 150). Strictly speaking, an SC plan should only contain a representation of the context being drawn and no other. However, in practice, many of the plans that appear on the back pages of Priniatikos Pyrgos context sheets include multiple contexts and so are in fact multi-phase plans.

135 The αποθήκη is the name given to Greek archaeological post-excavation centres. It literally translates to the English word ‘workshop’.
the site could be accounted for when deposited in the project’s storage facilities at the end of the day.\textsuperscript{136}

The recording of loci was never formalised to the same extent as was the case for contexts. Instead, information linked to the excavation of loci was recorded within the excavator’s site notebook. This more freeform approach to archaeological recording is encountered quite a lot in the archaeological field (Pavel 2012 pp. 6–7). Its roots go back to some of the earliest excavational practice (Harris 1989 p. 25). However, this is not to say that notebook recording is an anachronism in modern archaeology. Its use in the 21st century is usually a conscious decision, reflective of a project’s wish to be less dogmatic in terms of its recording practice (Hanson & Cheetham 2011; Pavel 2011). For these reasons, it is a popular tool of the post-processualist (Hodder 1997). The pail entity, which has been introduced above, was recorded on site using the same pro forma sheets used to record sub-contexts. As such, a pail’s paper correlate is primarily focussed on the cataloguing of the material found during its excavation.

\textsuperscript{136} Sub-context sheets were typically referred to by the project members as check-in sheets for this reason.
Sample sheets are another paper record type, which were employed at Priniatikos Pyrgos (Figure 84). These record the circumstances and the empirical observations made during the taking of an environmental material sample. Samples have become an increasingly important aspect of archaeological field practice as post-excavational scientific analyses become more accessible to the average project because of the falling costs of processing and increased archaeological expertise in this area (Coil et al. 2003; Turney et al. 2014 chap. 1). Following the example set by the context sheets, the sample sheets also include a space for the drawing of both a section and plan sketch. While these drawings are not intended to be highly detailed or overtly accurate in terms of their use of scale or their conformity to conventions, their use again emphasises the importance of spatial recording in the archaeological field method (Lynam 2013).
To supplement the plan and section drawings contained within the context and sample sheets, the project also employed the use of larger drafting sheets, on which plans, sections and elevations were drawn (Figure 85). Known as ‘supplementary drawing sheets’, typically, these records were completed by the project architect, although not always. The project also produced a sizeable number of field photographic records in keeping with modern archaeological common field practice (Dorrell 1994). Digital photography has democratised the simple creation of what can be highly detailed and informative image resources.137 While in the pre-digital age there was a need to employ specialist site photographers, now it is more common for each excavation area or even each archaeologist to be given their own digital camera and the control as to which and how much content is to be recorded photographically is placed in the hands of the excavator. While the increase in the taking of photographs during excavation can be welcomed generally as it provides an invaluable resource for the future interpretation of material and stratigraphy, this relatively recent practice has introduced a substantial new overhead in terms of record archiving (Stephen & Morgan 2014) as we will see in our

---

137 For some theoretical insights into the changes brought about by the introduction of digital photography into archaeology see Shanks (1997 p. 81) and Hamilakis et al. (2009).
discussion of the processing of the Priniatikos Pyrgos photographic record later in this chapter. On-site photographs and supplementary plans and section drawings were assigned unique identification numbers in a central site register at Priniatikos Pyrgos and these identifiers were subsequently used to associate these images with other records.

Records created in the field were also commonly changed during the post-excavation phase at the project. As part of their post-site daily routine, excavators would update their context and locus records and, occasionally, these records were also changed during the study seasons. Pottery readings, which constitute one of the more important and time-consuming aspects of the post-excavation archaeological process for most East Mediterranean projects (Wardle et al. 2001 chap. 7), were included as notes on the sub-context and pail sheets (Figure 83).

![Figure 86: the Priniatikos Pyrgos 14-7852 catalogue sheet](image)

The post-excavation phase at Priniatikos Pyrgos produced two other paper record types in the form of catalogued artefact sheets. The first was used to record the analysis of ceramic pottery and the second was used for catalogued items, such as glass, stone, ceramic non-pottery, bone and metal artefacts (Figure 86). These catalogue sheets were used to record the empirical measurements taken of the associated artefact and they also included the interpretation of the specialist who read the artefact. The cataloguing
process also saw the assignment of a type or more often a series of types to a particular artefact and we shall see how the observations made and decisions taken at this stage have filtered through the interpretive process, ultimately coming to influence the site’s narrative.

**Building an Archaeological Information System**

**Data gathering and archaeological schools of thought**

In the last section, we described the various entities used to record the archaeological processes undertaken at Priniatikos Pyrgos. These records were largely presented as vessels of objective truth. The majority of the documents discussed were of the pro forma type, meaning that they collected information in a systematic way with information of one type being entered into one specific field and information of another type being entered into another field. We define this type of recording as empirical, in that it documents information gleaned as a result of an actor making observations concerning a particular subject. Measurement fields are a canonical example of the output of this type of process; the archaeologist uses a tape measure to observe a context’s dimensions, which are then entered as numbers into the field assigned for this observation.

As was commented upon in our discussion of archaeological data in Chapter 1, a model which sees empirical data being derived from a series of observations, which ‘speak for themselves’ in a sort of atheoretical space, has since the introduction of postmodernist thinking into archaeology become problematic on a number of different levels. This was largely the model followed by the earliest archaeologists. Hodder and Hutson (2003 pp. 16–17) call this the ‘data ➔ theory’ model of knowledge, in which data precedes and indeed creates theory – essentially, this defines induction. The second basic framework for knowledge construction in archaeology is the ‘hypothesis testing’ model. In this, a series of logical propositions combine to form a hypothesis and these hypotheses can be interrogated logically in order to eventually propose theories. The idea that logic alone can be used to create idealised hypotheses is a central strut in the philosophy of logical positivism (Hempel & Oppenheim 1948) and positivism was employed extensively by the proponents of the New Archaeology movement in the 1960s. The ‘hypothesis testing’ model takes abstract hypotheses and tests them by feeding them with empirical

---

138 See Chapter 3 for more on both Aristotelian empiricism and logical positivism.
observations (Binford 1965). If they break, they are tweaked or discarded and if they remain intact, they are adopted until a better model is proposed.

The post-processualists reacted against this ‘processual’ framework by proposing that all observation is theory-dependent and so in the case of our context measurement example, the act of measurement and recording is dependent upon the parameters used to undertake and record the measurement. In other words, the post-processualists would argue that the idea of fact, or more accurately objective fact, is impossible to maintain. The battle between those archaeologists that believe that all interpretation is relative and those that work within a framework in which facts and even truth become obtainable has waged since the late 1970s and to an extent all archaeological interpretation from that point on exists either explicitly or implicitly within one camp or the other. The processualists argue that pragmatism must win out in order for the discipline to move forward and in order for it to focus on its primary job of work: to study the past (Binford 1982; Boudon 2005). While the post-processualists would counter that there is little point in working towards a goal that is premised upon a fundamentally flawed theoretical beginning (Shanks & Tilley 1993).

How then are we to reconcile these two opposing positions in the context of the recording of the Priniatikos Pyrgos material record? Considerations of archaeological theory can often be pushed to one side in discussions of data management. This view says that these are matters best left to the discipline’s theoreticians. However, choosing to side with one school or other is not simply an exercise in academic discourse; this decision consciously and subconsciously affects the ways in which the archaeologist performs almost every task that they encounter in their daily routine. It determines the parameters in which a trench is excavated, how a change in soil is interpreted, how artefacts encountered within that soil are read, how these observations are committed to record and how these records are ultimately represented and turned into knowledge. Taken together, these processes define archaeology, and, as such, they define the steps taken to reconstruct the past. They are all, therefore, meaningful and anything that affects them, be that an intellectual climate or otherwise, needs to be acknowledged and properly understood. It is, therefore, vitally important that before we talk of how the Priniatikos Pyrgos dataset was converted into RDF, the philosophical context of this series of practices be stated clearly.
On the one hand, the cleaning of the Priniatikos Pyrgos data and its alignment to the CIDOC CRM data model as it is represented using RDF are all processes that are consistent with archaeological processualism. Having said that, efforts were made throughout this process to be consciously aware of the decisions being taken. And this reflexivity is explicitly identified by Hodder (1997, 2003) and others (Chadwick 2003) as a central strut of postmodernist archaeological thinking. As such, this work exists within an intellectual framework that sees value in elements of the processualist and post-processualist schools of thought. This introduces a tension at the heart of the project’s method, and in many ways the remaining chapters all address particular aspects of this dialectic. At one extreme, there is the attraction towards the pragmatism that comes with the processualist approach, while at the other, the project constantly fights against the dangers of falling into the overly normative or deterministic strictures that processualism can promote (Shanks 2009). We return to this topic in Chapters 7 and 8.

Assessing the user needs at Priniatikos Pyrgos

During the period that Priniatikos Pyrgos was being subjected to archaeological scrutiny, a wide range of methodologies were applied to the study of its material record, and in many ways Priniatikos Pyrgos can be seen as a test plate for modern archaeological field practice given the breadth of its approach. The site and its environs have been investigated spatially by palaeomorphologists, geophysicists, GIS specialists and site architects. The excavated material record has in turn been processed by experts in ancient ceramics and lithics analysis, as well as by other specialist material type investigators. The organic remains have similarly been studied extensively by a combination of physical anthropologists, zooarchaeologists and experts in archaeobotanists. While there are commonalities of practice across all of these various categories of research activity, there are also as many methodological idiosyncrasies that need to be considered.

From a data perspective this translated into a wide range of demands being placed on the project Archaeological Information System. We have already seen how this was manifested in the form of the Priniatikos Pyrgos paper records but it also had implications for the transition to a digital information platform. In order to assess these

---

And this is ultimately where most archaeological research is situated.
demands, an effort was made throughout the building of the digital Archaeological Information System to engage with this community of users. In practice, this negotiation took place informally. No structured user forums were ever established to codify a list of user requirements. However, informal communications between the project research community and the Archaeological Information System developers regularly took place. This process evolved through time and the following sections are a record of that history.

The following is a summary of the high-level user requirements that came out of that process of engagement:

1. The Priniatikos Pyrgos digital resource structure needed to mirror that of its paper record.
2. The user interface design should be graphical, visually pleasing and, most importantly, intuitive.
3. Archaeology is essentially a matter of contextualisation. As such, the Priniatikos Pyrgos digital dataset needed to maintain the entity relationships identified during excavation and post-excavation.
4. These entities must be represented digitally as image, text and number data types.
5. The system must allow for more than one individual to input or access its resources simultaneously.
6. It must be accessible both during and in the inter-season period.
7. It needed to be accompanied by a robust archiving regime, ensuring the sustainability of its data in the short to medium terms.

The first digitization phase

From as early as 2005 it was understood that the project’s paper record would need to be digitised and a FileMaker database system was selected to satisfy this need. FileMaker is a database system that is used by many archaeological projects (Motz & Carrier 2013; Wallrodt 2012). It has a number of advantages. It is graphical, it allows for the management of multiple data types and it enables more than one user to work on a single database project at the same time. But it also has disadvantages. First, it is proprietary or closed software and, therefore, constrained in terms of its extensibility and openness. It is also costly and this can present a severe limitation for many archaeological projects given the current embattled state of archaeological funding models (Schlanger et al. 2010).
In 2009, the database was fundamentally redesigned and this also heralded the beginning of the large-scale digitisation and inputting of paper records. In total, this digitisation effort has cost the project in the region of 3,872 hours of labour since 2009.\textsuperscript{140} Currently\textsuperscript{141}, the FileMaker database is 6.24GB in file size and it contains 692 context, 1,719 sub-context, 341 locus, 1,332 pail, 3,058 catalogued ceramic pottery, 2,384 catalogued artefact, 668 catalogue drawing, 1,978 catalogue photograph, 4,362 field photograph and 1,812 scanned paper records (Figure 87).\textsuperscript{142}

![Figure 87: a bubble chart showing the relative sizes of the various Priniatikos Pyrgos FileMaker tables](image)

\textsuperscript{140} This calculation is based on two people working full-time for 11 hours per day, 5.5 days per week for 6 weeks per season during the 2009-10 campaigns and 5 weeks per season during the 2011-14 study seasons. It is often argued that the high costs involved in the introduction of digital tools use in field archaeology rule them out as a viable possibility for most archaeological projects. However, looking at the scale of the labour costs accruing as a result of digital practice not being implemented at an early stage in the archaeological knowledge creation process at Priniatikos Pyrgos, this strategy of opting for labour over technological investment might well be short-sighted and indeed a false economy (Austin 2014; Butina 2012; Ellis & Wallrodt 2011a).

\textsuperscript{141} As of 1 June 2015.

\textsuperscript{142} In 2010 it was decided to create a scan image of each paper record and to enter these images as records in a paper records table into the FileMaker database. This work was carried out as a protection against possible accidental loss or damage of the project’s paper archive. See Aitchison (2004) for more on the hazards of archaeological storage, paper or otherwise.
Migrating to the Archaeological Semantic Web

In 2010 a decision was made to move away from the project’s total dependence on the FileMaker system, towards the use of an Open-source Software and Open Data Archaeological Information System. While FileMaker had certainly fulfilled its function in the sense that it had allowed the project to install a digital database platform quickly, it soon became apparent that the system’s lack of software and data openness was going to present problems in the future. Research archaeology can be very different from other forms of humanities research in the sense that its practitioners often have highly limited and time-specific access to their source material, often no more than a handful of weeks for every year. This means that archaeologists need to rely on their records and increasingly the digital material record is filling this access void. If you add to this scenario the fact that archaeological projects tend to be made up of multinational teams, who must return to their home institutions at the end of each field season, the advantages of employing digital media become obvious. As such, the Priniatikos Pyrgos project needed to be able to publish its raw data to the web to allow for its further study and as part of a more general dissemination strategy. It was also felt that the FileMaker system was too isolated from other archaeological datasets. While there exist a number of hosting platforms that allow for the online publishing of FileMaker data, these solutions tend to be prohibitively expensive, especially when the archaeological dataset in question needs to be maintained indefinitely. There is also the matter that this published data would be presented in a closed and proprietary format to a limited user group.

Around this time, I attended the 40th Annual Conference of Computer Applications and Quantitative Methods in Archaeology (Earl et al. 2013), held that year in the University of Southampton. This introduced to me the subjects of Open Data, Linked Data and the Archaeological Semantic Web for the first time. However, despite the many compelling reasons for employing Linked Open Data as an Archaeological Information System for the Priniatikos Pyrgos dataset, I found that for the most part very few real archaeological datasets were using such techniques and there were even fewer ‘off-the-shelf’ solutions.


144 Data sustainability has long been an issue within archaeology. Archives, regardless of their type, are made up of perishable materials and curators are constantly strategizing on ways of alleviating this working challenge. If anything, the move towards digital storage in recent times has exacerbated the problem with many viewing digital media as even more susceptible to future access problems than is the case for paper resources (Kansa 2011 pp. 18–19; Rains 2011 p. 166).
available for use with archaeological data. After a number of months of further investigation, I decided that the best option would be to build a Linked Open Data solution and the linkedarc.net system ultimately came out of this process.

**Choosing an ontology for the mapping**

**linkedarc.net resource**

Additional notes on the linkedarc.net custom ontologies

http://linkedarc.net/ontologies/la

Before we talk about the mapping process itself, we need to consider the subject of the ontology that the Priniatikos Pyrgos dataset will be mapped to. The benefits accruing from the use of public ontologies have been explained in Chapter 3. This account also introduced the CIDOC CRM and CRM-EH ontologies in the context of their use as a data model for cultural heritage and archaeological datasets. Chapter 3 also explained how SKOS has become by far the most popular ontology on which to build controlled vocabularies. The lion’s share of the Priniatikos Pyrgos ontological needs is delivered using these three foundational ontologies. Other models appear in certain contexts (for instance, FOAF and the SCHEMA models are used in part to model aspects of the social data of the project) but the CRM, CRM-EH and SKOS predominate. There is no need to repeat the outlines of these models already presented in Chapter 3. The following accounts will help, however, to flesh out the concepts introduced in Chapter 3 using real archaeological data scenarios.

Public ontologies are not, however, the only data models used by the project. We have already commented on the uniqueness of certain entities employed as part of the Priniatikos Pyrgos paper record. The sub-context is an obvious example but it was also the case that a number of the fields within the FileMaker database needed proprietary handling. To provide for both of these scenarios, it was necessary to build a custom ontology.
Before dealing with the Priniatikos Pyrgos ontology, we need to first introduce the linkedarc.net system ontology on which the former is built. This ontology, which has the URI base <http://linkedarc.net/ontology>, exposes a number of classes, which all derive ultimately from the CRM’s E70_Thing class (Figure 88). It is a relatively simple model, whose sole purpose is to organise linkedarc.net project data, whose structure was introduced in Chapter 4. The la:Project class models linkedarc.net projects, while the la:Datatable and la:Record classes model datatable and data record concepts respectively. The ontology also includes the la:ArchaeologyThing class, which represents a generic archaeological entity. It is intended that the la:ArchaeologyThing class be extended by project-specific ontologies, which wish to model custom classes related to the field of archaeology.

The system ontology provides a number of new predicates, which are linked to these classes and similarly relate to the matter of structuring linkedarc.net project data. The la:records-ontology-classname is used by la:Datatable entities to point to the entity class, which they are associated with. The la:project predicate is used extensively by the Priniatikos Pyrgos mapping. It allows any entity to be linked to a particular linkedarc.net project. This is useful when querying linkedarc.net data, as it allows the query to first specify that only data of a certain project be considered. The la:ontology-url predicate allows an la:Project entity to be linked to the URL of its custom ontology file and the la:uri-base predicate specifies the base URI address where all la:Record entities within an la:Datatable entity reside.
While somewhat larger and more diffuse than the linkedarc.net system ontology, the semantics of the Priniatikos Pyrgos project ontology are still relatively straightforward. Its URI base is <http://linkedarc.net/ontology/la_pp> and Figure 89 describes its hierarchy of classes. While all of the linkedarc.net system ontology classes derive from the la:Thing class, the Priniatikos Pyrgos hierarchy is less well ordered with about a half of its classes originating from the la:Thing class and the other half deriving from specific CRM or CRM-EH classes.\footnote{While the internal hierarchy of classes is not shown in this diagram, all CRM and CRM-EH classes ultimately derive from the CRM E1_CRM_Entity class.}
Remember that this ontology was built to provide the semantic functionality required to adequately model elements of the Priniatikos Pyrgos dataset, which are not accommodated by the CRM and CRM-EH models. The Priniatikos Pyrgos sub-context administrative unit is provided for by the la_pp:LA_E15_Subcontext. In order to differentiate between SKOS entities dealing with period data on the one hand and more general vocabulary information (vessel types, colours, et cetera) on the other, all entities of the former category were also assigned the class la_pp:LA_E19_Period (inheriting from ecrm:E52_Time-Span) as a type, while the latter were given the type la_pp:LA_E20_VocabularyItem.

In order to model the locus and pail data created by the project, it was necessary to create the la_pp:LA_E12_Locus and la_pp:LA_E13_Pail classes and these both derive from the la_pp:LA_E11_AmericanRecording class, which in turn is a subclass of the la_pp:LA_E10_RecordingMethod class. For completeness, la_pp:LA_E10_RecordingMethod class also parents la_pp:LA_E14_MoLASRecording, which represents the MoLAS recording method concept. Neither the la_pp:LA_E11_AmericanRecording and la_pp:LA_E14_MoLASRecording classes are instantiated as entities by the mapping process.

The project image data was modelled using the la_pp:LA_E18_Image class, which inherits from ecrm:E38_Image. Trenches were also accommodated using an extension class, la_pp:LA_E3_Trench, which extends the crmeh:EHE0004_SiteSubDivision class. The final set of Priniatikos Pyrgos concepts that needed to be accommodated by extensions were related to the catalogued items. As noted above, items were first classified as either being ceramic or non-ceramic objects. As the project evolved and material-specific specialists entered the project’s workforce, certain new categories of non-ceramics object record were introduced into the project’s pro forma record set. For example, the project’s substantial glass collection underwent a thorough examination by a glass specialist in the 2009 season and this resulted in the creation of a whole range of 146 The syntax used in the naming of these new classes and predicates largely follows that of the CRM and the CRM-EH. All classes are labeled in the form LA_Exx_yyyyyy. The ‘LA’ prefix indicates that the class is a part of the linkedarc.net knowledge domain. ‘E’ stands for entity and ‘xx’ is the index of that class in question. ‘yyyyyy’ is a descriptive name for the class, following the normal class capitalization conventions. Predicates follow the form ‘Pxx_yyyyyy’, where ‘P’ indicates that this is a URI for a predicate, ‘xx’ is the index of that predicate and ‘yyyyyy’ is its descriptive label.
new fields in the FileMaker database to accommodate this new recording practice. The same phenomenon took place for the loomweight and spindle whorl assemblages in the 2013 season. It would have been entirely possible to include all of this content (catalogued ceramic and all non-ceramic finds) as crmeh:EHE0009_ContextFind but for procedural reasons it was decided to create a small hierarchy of classes (la_pp:LA_E5_Catalogue, la_pp:LA_E6_CataloguedObject, la_pp:LA_E7_CataloguedGlass, la_pp:LA_E8_CataloguedLoomweight and la_pp:LA_E9_CataloguedCeramic) to define these catalogued item groupings. Each of these catalogue classes ultimately derives from the crmeh:EHE0009_ContextFind class.

As in the case of the linkedarc.net system ontology, the Priniatikos Pyrgos ontology does not just provide new classes; it also makes available a number of custom properties to be used as predicates alongside these new classes and existing CRM and CRM-EH classes. The vast majority of the 38 properties included in this collection are designed to be used by crmeh:EHE0007_Context entities. For example, a la_pp:P01_has_subcontext predicate allows for the linking of crmeh:EHE0007_Context entities with instances of the la_pp:LA_E15_Subcontext class and the la_pp:P50_images_contexts property associates la_pp:LA_E18_Image and crmeh:EHE0007_Context entities. However, other classes are extended as well, with the la_pp:LA_E1_ArchaeologyThing, la_pp:LA_E12_Locus, la_pp:LA_E18_Image and la_pp:LA_E13_Pail all having particular predicates added.147

**Mapping the Priniatikos Pyrgos dataset to linkedarc.net**

Once the decision was made to move away from FileMaker, the linkedarc.net development process began and this is described in Chapter 4. In tandem with this, an equally important plan was put in place to transform the FileMaker data so that it could then be imported into the evolving linkedarc.net system. Conveniently, FileMaker Pro allows the user to export to a number of different formats: Excel, XML, TSV and CSV. The challenge of this data-mapping phase of the project was to develop a workflow that would be capable of converting this data into RDF form, which could then be inputted into the linkedarc.net server.

147 For a complete definition of the Priniatikos Pyrgos ontology, visit http://linkedarc.net/ontology/la_pp.
Figure 90 describes the high level workflow used to map the data. FileMaker, as we know, is an RDBMS database provider. This means that its data is constituted of rows and columns within a series of tables. The Priniatikos Pyrgos FileMaker database contained a number of different tables, one for each paper record type already introduced. Each of these tables was exported from FileMaker Pro as a CSV file. Each CSV file was then imported as a new project into the Open Refine application. Open Refine was used to first clean the data and then map it onto an RDF output using the RDF Refine extension. The RDF output was then imported into the linkedarc.net Apache Jena triplestore.

We will now consider in detail the cleaning and mapping of data using the Open Refine application.

**Messy data and data field types**

Before we talk of data cleaning and mapping, we must ask exactly what it is that we mean by ‘messy’ data? Perhaps it is best to first consider what it is not. Schöch contrasts messy data with what he refers to as ‘smart’ data, which is data that is ‘structured or semi-structured… contain[ing] markup, annotations and metadata’ (Schöch 2013). We have already introduced the difference between structured and unstructured data in Chapter 1.

An example of structured data is a spreadsheet document and an image is an example of...
unstructured data. The spreadsheet contains a prescriptive and, therefore, queryable data structure. We can ask questions of parts of it. It need not be simply viewed as a monolithic block of meaning. The image on the other hand is unstructured. Excluding the use of advanced image processing techniques, parts of an image cannot be interrogated as independent entities in their own right.

Figure 91: the data structure spectrum

Following Schöch, messy data is, therefore, unstructured but there is no simple binary separation between structured and unstructured data. Instead, the structure of any one dataset is measured on a spectrum with highly structured data at one end and entirely unstructured data at the other (Figure 91). In practice, these extremes are never inhabited by any actual data format.

In our context, the messy data coming out of the Priniatikos Pyrgos dataset was structured in the sense that each node occupied a cell within a relational database table. However, when each cell was viewed in the context of the same cells in other records, there was found to be significant variations in the format of the data. And the main reason for this was that the majority of the FileMaker fields used in the Priniatikos Pyrgos data set were declared as free-text fields. This means that the person inputting the data into these fields is not constrained by what they are allowed to enter. As a result, fields that deal with concepts, which in hindsight could or should have been constrained (for example, Munsell measurements, object dimensions, et cetera) came to be populated with free-text entries.

The question of whether a database field should be made as flexible as possible or whether it should be limited in terms of what a user can enter into it, is not simply a methodological matter. It has much wider philosophical implications and these were touched upon in our discussion of how archaeological schools of thought impinge on archaeological data recording practice. If the data designer decides to make a field as
free-text, then it could be argued that this is more compatible with post-processualism than it is with processualism; in the sense that, the amount of constraint being placed on the user is minimal,\footnote{There will always be some constraints placed on the user regardless of the system used. For instance, a text field can accept only text. The user cannot decide to enter an image into such a field. And text fields must be serialised using language and language, while enormously flexible, is still a system and a system that imposes its own constraints (Wolf 2015).} which in theory should accommodate a wider range of interpretations.

If, on the other hand, the data designer were to make the user select from a drop-down list of entries when filling in a particular field, then there is less sense that the user is entirely in control of the selection. Yes, they are allowed to choose from a list of values but this set is finite and is ultimately determined by the system’s creator and not the user. This second approach is very practical in the sense that it makes the affected data more indexable by future users but it needs to be borne in mind and understood that such pragmatic design decisions do have philosophical implications down the line; in this case, it makes the dataset more aligned with processual ways of thinking.

**Data cleaning using Open Refine**

The cleaning of data is fundamental to any data mapping or data aggregation process. Given the recent rise in the creation of unstructured data, such as words, images and video in both public and private domains – under the umbrella term of Big Data (Lohr 2012), that ability to efficiently clean a dataset has become increasingly important (Hernández & Stolfo 1998). While the data cleaner is certainly not limited in terms of the tools that they can use to carry out this process\footnote{Excel is another popular application used to clean data and at the beginning of the Priniatikos Pyrgos linkedarc.net project, it was used to perform the lion’s share of this type of processing.}, the Open Refine application\footnote{Open Refine installs locally on a Windows, Mac OSX or Linux operating system. In an era of web-based applications, the fact that Open Refine is run locally is somewhat surprising but this apparent anomaly is explained away as an intentional feature of the system. For an application that processes large amounts of data, some or all of which might be private in nature, it makes sense that the data processing should be carried out entirely in a local context.} stands out for a number of reasons (van Hooland et al. 2013; Padilla 2015). First, it can handle very large datasets and while, in comparison to some other archaeological datasets, the Priniatikos Pyrgos dataset is not enormous by any means, the ability to scale up in the future is an important capability to have. Open Refine can record and replay the steps taken during a mapping process, saving the user a substantial amount of time in scenarios in which data needs to be reimported from source and we talk about this
Chapter 5 - Populating linkedarc.net

feature more at the end of this chapter. It has scripting support and it also provides a number of different export formats and the ability to install extensions, which add to the application’s baseline functionality.

Figure 92 shows a screenshot of the Open Refine system having just imported the pails table from FileMaker Pro. Open Refine allows you to view the range of values contained within a specific field. This feature enables the user to quickly spot and correct potentially troublesome data. In this case, the range of values entered into the ‘excavators’ fields for all pail records is summarised on the left-hand side of the screen. Open Refine allows you to pick out systematic errors (those being errors that are repeated consistently over a section of the values) and errors arising as a result of unconscious human behaviour, such as spelling mistakes or the inclusion of line breaks within a value where there should be none. The sorts of variation seen in these values have come about as a result of the laissez-faire approach to data structuration that free-text fields allow. As we will see, the cleaning up of this type of data can be time-consuming and it is a practical demonstration of how basic data design decisions, whether they are informed explicitly by conscious philosophical decision making or not, can have significant knock-on affects down the line.

Open Refine supports a scripting language known as OpenRefine Expression Language or GREL. This is a subset of the JavaScript language and while relatively rudimentary, it

---

156 Note that the version of Open Refine used during this project retained the Google Refine iconography.
is flexible enough to allow for the writing of scripts, which can be used to target such data errors as is evidenced in Figure 92. An example of a GREL script is reproduced below. Running a script on a data column effectively repeats the scripts logic for each value contained within the column. The script shown in Table 19 replaces any forward slash characters found within a value with a comma character.

```
1. value.replace('/', ',');
```

<table>
<thead>
<tr>
<th>Table 19: using the GREP replace function</th>
</tr>
</thead>
</table>

The corrections needed to resolve the types of errors encountered in the Priniatikos Pyrgos FileMaker tables were numerous and varied and depended on the form of the data coming in and the intended form of the output. Imagine that we want to clean up the values found in the ‘excavators’ field shown in Figure 92. The output that we want from this process is a series of URLs in canonical form such as <http://example.org/excavator-sdb>. Some of the input values contain demarcator characters (‘/’, ‘and’, ‘,’), which separate a number of different excavators contained within the one text value. To handle this, we first standardise the demarcators by converting them all to the same character, changing them all to be commas, for example. We then look for these commas and create a series of new columns, one for each new value. Thankfully, Open Refine includes a function to do just that and by running ‘Split into several columns’, we end with a series of new columns called ‘excavators1’, ‘excavators2’ and so on (Figure 93). The point here is that there is no standard approach to data cleaning. Every new problem will require a slightly different solution. The solutions documented in this account are included to illustrate the types of functions that Open Refine is capable of. If you wished to clean a dataset other than the Priniatikos Pyrgos Project dataset, this would undoubtedly demand the employment of a different set of strategies.

![Figure 93: the results of processing the 'Excavator(s)' field in the contexts table](image)
Calendar dates are an important value type being employed extensively within the Priniatikos Pyrgos dataset and they are also a type of information that is often prone to inconsistency (Rahm & Do 2000). They were used to record the dates of excavation, when a sheet was first created, when it was completed, when catalogued objects were processed and so on. In order to map dates to RDF, they need to be in the format YYYY-MM-DD, which the majority of the Priniatikos Pyrgos date values were not. In order to format these values correctly, more GREL scripting was needed as is shown in Table 20.

```
1 value.toDate().datePart('years')+'-
2 '+'with(value.toDate().datePart('months'), v, if(v.length()==1,
3 '0'+v, v)) +'-'+with(value.toDate().datePart('days'), v,
4 if(v.length()==1, '0'+v, v))
```

**Table 20:** cleaning dates using GREL

In this script, the value in question is converted into an Open Refine Date type, before its constituent parts are recombined into the required YYYY-MM-DD form. The result of this operation is shown in Figure 94.

![Figure 94: converting messy date values into RDF date format using Open Refine](image-url)
Processing number values using RDF Data Utilities

Fields containing number (integer or floating point) values are particularly prone to errors as a result of the inconsistent use of measurement units. Let us consider the processing of a field called ‘fine-grained pottery weight’ in the pails table. The values in this field will be integers or floating-point numbers. They represent the weight values of the fine-grained pottery found during the excavation of each pail. The problem emerges when one set of values uses one unit measurement (for example, grams) while another uses a different unit measurement (for example, kilos). Resolving whether, in the absence of an accompanying unit label, if a value of ‘1.100’ equates to 1.1gr or 1.1kg can be tricky and deciding to choose one unit over the other ultimately comes down to probability.

![Figure 95: the distribution of fine-grade pottery weights across all Priniatikos Pyrgos pail records](image)

One solution to this dilemma is to analyse the entire range of values and to consider its general pattern as is illustrated in Figure 95 (Hellerstein 2008). It also helps enormously if the subject of the values is understood by the individual carrying out the data mapping. For example, if the majority of the values are relatively low (≤ 10 for instance) as is shown in Figure 95 and we know that these values are used to measure the weight of small ceramic objects, then it is probable that the ‘1.100’ value referred to above equates to 1.1gr and not 1.1kg. This example clearly highlights the importance of domain knowledge in the process, which the humanist brings, and reminds us that data cleaning...

---

157 Each Priniatikos Pyrgos pottery unit was first washed, dried and then sorted into fine, medium and coarse-grained collections. These were then weighed and counted and this information was recorded on the corresponding pail or sub-context sheet.

158 Note that while some of the ‘fine-grained pottery weight’ values were greater than 10, they are very few in number relative to those values that are less than or equal to 10 and as a result difficult to see in this chart.
is an inherently interpretive act. While large arrays of values can be, and typically are, regularly processed using algorithms, the parameters that define the operation of these automated processes are determined by the intuition, reason and experience of the human agent.

In order to systematically deal with these sorts of number unit problems, I designed and built the RDF Data Utilities web app. This web app allows the user to align a range of number values to the one unit scale. The user selects a CSV file, which contains the potentially inconsistent data. She then selects a specific column to analyse. RDF Data Utilities then processes the range of values and presents a summary of this analysis to the user (Figure 96). The user can then use their knowledge of the dataset and their common sense to decide how to best align the data range.

Figure 96: the RDF Data Utilities application displaying a summary of the contents of a number column

159 http://rdfdatautils.linkedarc.net
Aligning a series of number values in this case means to format each value so that it conforms to a particular scale. For our example, all of these weight values should have been entered as grams. The RDF Data Utilities wizard aligns the full range of weight values to the correct unit (Figure 97) and then outputs the updated data as a CSV file, which can be imported back into Open Refine for additional processing. In the case of the pails table’s fine-grade pottery weight range of values, the RDF Data Utilities application changed the scale of 40% of the values. The operation takes about a minute to complete, which is a fraction of the time that would be needed if the 1,332 pails records were to be processed by hand.

The alignment of a range of numbers as is described above highlights how data cleaning involving thousands or even millions of values will inevitably introduce occasional errors into a dataset (Hellerstein 2008 p. 2). The pragmatist would argue that once cleaned, the data becomes exponentially more useful as a scholarly resource than its messy predecessor. Take the example of the fine-grained pottery weights referred to above. Once aligned, the researcher is able to analyse the pails dataset in the context of what it says about changing patterns of fine pottery use across time and space. This is meaningful information and can be used to construct more complex interpretations. The pragmatist would point out that, were the data’s units to remain inconsistently applied, that any analysis attempted on the dataset would present erroneous and misleading results. The irony of course of this last statement is that the inevitable introduction of errors into a dataset as a function of its being cleaned presents the same danger. This matter of changing an archaeological project’s raw data using automated means is clearly a huge issue and, ultimately, it is one for which each individual project must form an opinion. Following best practice, each project should also ensure that their policy be rendered as transparent as possible to the wider community. For the Priniatikos Pyrgos dataset, the decision was made to proceed along the pragmatist’s route by using
automated data cleaning and, therefore, to accept that occasional errors will be introduced into the data.¹⁶⁰

**Processing periods using RDF Data Utilities**

The RDF Data Utilities tool also provides a value mapping function, which can be used to map controlled vocabulary candidate data, particularly period data. Periods are a notoriously challenging value type for the ancient world scholar to deal with (Binding 2010; Rabinowitz 2014). We have already discussed the work of the PeriodO project in Chapter 4 and while this initiative does not explicitly deal with the problem of aligning messy period data, its existence does highlight just how important period data processing is to the successful workings of most Archaeological Semantic Web resources.

For the Priniatikos Pyrgos project, period values occur in pail, sub-context and catalogued object and ceramic records. Each of these types of record is at some point assigned a period range by one or more specialists. For example, let us imagine that the sub-context 001-001’s pottery is read by a number of pottery specialists and they assign it a date range of ‘Hellenistic to Roman’ based on the occurrence of diagnostic sherds of these periods within its assemblage. This is written down on the sub-context sheet and eventually it is entered into the FileMaker database as free-text. A computer algorithm might be able to interpret this quite easily, if all of the sub-contexts deemed to be of this date were also assigned the ‘Hellenistic to Roman’ text value. Unfortunately, more often than not, the same conceptual value can be and is verbalised in many different ways. For example, another sub-context read with the same date range might be entered as ‘4th C BCE to 3rd C CE’ or using a combination of centuries and period labels, as in ‘4th century BC to Roman’.

Obviously, the range of possibilities for dates is wide. Ideally, you would convert these descriptive text values into a start and an end period, which can then be used to create a list of all relevant periods. For the example, the ‘Hellenistic to Roman’ value could be converted to an array containing the ‘Hellenistic’, ‘Early Roman’ and Late Roman’ periods. The most dependable way of achieving this is to use human agents to review and update each record one after another. In most data mapping exercises, however, this would be impractical. And so, for the most part, an automated component must play

¹⁶⁰ How to monitor the introduction of these errors is a separate and difficult question. In practice, many archaeological projects adopt a randomised human-based monitoring system.
some part in the solution. The value mapping function of the RDF Data Utilities tool was created to deliver this type of automated processing.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>from</td>
<td>to</td>
<td>index</td>
</tr>
<tr>
<td>'prehistoric'</td>
<td>[fn, 'emi', 'emii', 'emii', 'mmi', 'mmii', 'mmiii', 'lmi', 'lmi', 'lmi']</td>
<td>0</td>
</tr>
<tr>
<td>['fn', 'neo', 'finalneolithic']</td>
<td>fn</td>
<td>0</td>
</tr>
<tr>
<td>'prehistoric'</td>
<td>[emii, 'emii', 'emii', 'mmi']</td>
<td>1</td>
</tr>
<tr>
<td>'prehistoric'</td>
<td>[emii, 'emii']</td>
<td>1</td>
</tr>
<tr>
<td>'emii', 'emii'</td>
<td>'emii'</td>
<td>2</td>
</tr>
<tr>
<td>'emii', 'emii'</td>
<td>'emii', 'mmi']</td>
<td>3</td>
</tr>
<tr>
<td>'emii'</td>
<td>'emii'</td>
<td>3</td>
</tr>
<tr>
<td>['emii', 'earlyminoan']</td>
<td>'emii'</td>
<td>4</td>
</tr>
<tr>
<td>['mmi', 'middleminoan']</td>
<td>'mmi', 'mmii']</td>
<td>4</td>
</tr>
<tr>
<td>'protopalatial'</td>
<td>'mmi', 'mmii']</td>
<td>4</td>
</tr>
<tr>
<td>'mmi'</td>
<td>'mmi']</td>
<td>5</td>
</tr>
<tr>
<td>'neopalatial'</td>
<td>['mmiii', 'lmi']</td>
<td>6</td>
</tr>
<tr>
<td>'neopalatial'</td>
<td>'mmiii', 'lmi']</td>
<td>6</td>
</tr>
<tr>
<td>['mmi', 'lateminoan']</td>
<td>'lmi']</td>
<td>7</td>
</tr>
<tr>
<td>['lmi', 'lmi']</td>
<td>'lmi']</td>
<td>7</td>
</tr>
<tr>
<td>'finalpalatial'</td>
<td>['lmi', 'lmiili']</td>
<td>8</td>
</tr>
<tr>
<td>'lmi'</td>
<td>'lmiili']</td>
<td>8</td>
</tr>
<tr>
<td>'postpalatial'</td>
<td>'lmiili']</td>
<td>9</td>
</tr>
</tbody>
</table>

Figure 98: a sample RDF Data Utilities value mapping plan for periods

The value mapping function works as follows. The user begins by selecting a CSV file to process. As with the number alignment function, they select a column, which contains the period values. They must then choose a rule set with which to conduct the mappings. They can select to use an in-built system mapping rule set, which has been designed to map period data or they can choose to use their own custom mapping file. A sample mapping file is shown in Figure 98. It contains a series of entries, each with a mapping ‘from’ and ‘to’ value and an ‘index’ value. The ‘from’ value is a series of words, which will be searched for by the algorithm when the mapping is run. If a match is found, the algorithm then creates the range of values, which are contained in the corresponding ‘to’ field. The index for the mapping is important as it tells the algorithm how each mapping is placed within an overall mapping sequence. This sequence feature might not be important for certain data types such as colours or artefact typologies for which order is not important. However for periods in which any one period is defined relative to preceding and subsequent periods, knowing the place of a period in its sequence is vital. If the algorithm is given an input such as ‘Prepalatial to Neopalatial’, it can use this index information to construct a range of periods that satisfies this input. Following this logic, ‘Prepalatial to Neopalatial’ would map to ‘emi, emii, emiii, mmi, mmii, mmiii, lmi’.

The periods contained within this mapping list was compiled by first manually sorting through all of the periods referenced by Priniatikos Pyrgos records of various different

---

161 This is the in-built system mapping file.
types. This list was then used to create a minimum set of periods that satisfies all of the records. The periods also needed to be associated with absolute date ranges. These were sourced in a number of standard textbooks dealing with the material culture of the Aegean region (E. H. Cline 2012; Dickinson 2006; Pomeroy et al. 2004). The lead project excavators and post-exvation specialists were also consulted in the drawing up of these absolute date ranges for the periods. The final list of periods is reproduced in Figure 99.

Once the mapping process is complete, RDF Data Utilities displays the updated data in table form to the user. The original set of column values is augmented with a series of new columns, which contain the range of periods induced by the system. The updated CSV file can then be downloaded and inputted back into Open Refine for further processing, if needs be.

Creating vocabulary lists out of messy data

Controlled vocabularies are a key instrument in the Archaeological Semantic Web model. We have discussed their functional justifications and philosophical implications at length in Chapters 3 and 4. In a nutshell, they vastly increase the indexability of an individual dataset and they allow for the interrogation of multiple datasets, which use the same or linked controlled vocabularies. For all of these reasons, it was imperative that suitable data from the Priniatikos Pyrgos be mapped to controlled vocabulary values. Again, this was a complex and tedious procedure but was helped by Open Refine’s flexibility. Again, human user oversight proved indispensible in the resolution of problems, such as dealing

---

162 As is described in Chapter 3, concepts in the same or separate vocabularies can be associated with one another within conceptual hierarchies using the SKOS, OWL and RDFS ontology predicates such as skos:exactMatch, owl:sameAs, owl:equivalentClass, skos:inScheme, skos:broader and rdfs:subClassOf.
with value synonyms (should ‘terracotta’, ‘ceramic’ and ‘plastic’ be considered to be the same concept or does it depend on the context of their use?) and ambiguities (does a ‘mortar’ value refer to a grindstone implement or a bonding material used in construction?). A purely computer-based mapping solution would find it very difficult to achieve this level of oversight.

Figure 100: Priniatikos Pyrgos FileMaker fields mapping onto vocabulary values

For any vocabulary mapping project, the first step that needs to be taken is to identify the fields in the source dataset, which would most benefit from and be most compatible with this type of mapping. Fields that contain conceptual information, such as context and artefact types, periods and agent roles are all potential candidates. Most Priniatikos Pyrgos tables held more than one suitable field (Figure 100). The catalogued ceramic pottery table, for instance, contained chronology, decoration, fabric, vessel type and ware fields, which could all be mapped to controlled vocabularies. A now familiar set of challenges presented themselves, mostly associated with the handling of free-text field processing. Inconsistent use of textual conventions and misspellings were widespread.
The second vitally important step was to choose a suitable controlled vocabulary for the dataset. Ideally, one would use an existing public ontology for the reasons (saving time, conforming to existing domain conventions) already outlined in Chapter 3. However, for various reasons (most notably, ignorance of best practice), it was decided at the time to create a custom vocabulary for the Priniatikos Pyrgos data, which was based on an aggregation of the concepts derived from the parsing of the values contained within the fields listed in Figure 100. While in hindsight this does not represent best practice, it is, nonetheless, a strategy adopted by many cultural heritage content providers. The principal reason for this is that these institutions are reluctant to relinquish control of such an important intellectual property as the vocabularies that they use to define their artefacts. Despite this failure of project planning, we will see that it is possible to retrospectively link custom concepts to public domain concepts using SKOS, OWL and RDFS predicates, which provides a workable middle-ground.

![Priniatikos Pyrgos Custom Vocabulary](image)

**Figure 101: the Priniatikos Pyrgos custom vocabulary value groupings**

At this point in the creation of the custom Priniatikos Pyrgos controlled vocabulary, it was necessary to draw up a rough plan of the categories of concepts that would need to be created in order to satisfy all of the Priniatikos Pyrgos mapping needs (Figure 101). It was also necessary to consider how, if at all, potential mapping conflicts, such as the use of synonyms, would be dealt with. These broad conceptual groupings either mirrored the
fields from which they came (for example, the ‘context type’ concept grouping was populated with values originating from the ‘context type’ field) or particular subsets of the source field (for example, the ‘catalogued ceramic pottery decoration’ field created various different groupings such as colour concepts and particular categories of decoration).

Next, the data needed to be mapped onto the concepts in the chosen vocabulary. There exist sophisticated Natural Language Processing techniques to achieve this type of concept extraction from text inputs (May et al. 2015) but in the interests of simplicity, a more rudimentary method was employed instead (Figure 102). Using Open Refine’s scripting capability, each relevant field was systematically searched for a particular text value. When found, a new column was created and populated with the corresponding vocabulary value. This approach demanded an extremely hands-on approach and it was selected primarily because an automatic process could not be trusted for the mapping of such an important aspect of the project data. Therefore, while this process proved time-consuming, this was justified as a necessary expense. In total, this processing, carried out
across all of the Priniatikos Pyrgos fields identified for vocabulary mapping, created a list of 258 unique concepts.

As already mentioned, custom vocabularies become infinitely more valuable when their component parts are linked to public vocabularies. In order to create these conceptual links, it was necessary to employ Open Refine’s web service functionality. This allows you to make calls to web services based on the contents of a particular field. This makes it possible to construct SPARQL queries that can be sent to the British Museum and Seneschal SPARQL endpoints. In so doing, it was possible to associate the Priniatikos Pyrgos custom vocabulary entries with links to concept URIs used by the British Museum, English Heritage, Royal Commission on Ancient & Historical Monuments of Scotland and Royal Commission on Ancient & Historic Monuments Wales.

Finally, any newly constructed vocabulary needs to be extensively documented using so-called ‘scope notes’. These values describe the meaning of each vocabulary concept and so are vital in terms of promoting a vocabulary’s re-use. A number of the concepts that came to populate the Priniatikos Pyrgos vocabulary were terms that were not solely related to archaeology. In order to remove some of the burden of manually entering all 258 scope notes, Open Refine’s web service call functionality was again used to ask DBpedia SPARQL interface for descriptions of concepts such as obsidian, wall and pumice. All of the descriptions gathered were validated manually.

**Geocoding the context and locus data**

Geocoding is the process of assigning geographic coordinates in the form of a longitude, latitude and elevation value to a location on earth (Goldberg et al. 2007). Geographical Information Systems (GIS) have had an enormous influence on the way that archaeologists conduct their business and interpret their data (Bevan & Conolly 2004; Conolly & Lake 2006). As we have seen with our description of the Single Context recording method, the archaeological method is primarily based on visual and, more specifically, spatial information. In that light, GIS and archaeology are a natural fit.

---

163 The Priniatikos Pyrgos custom vocabulary list is included as a linkedarc.net table and can be viewed at http://linkedarc.net/projects/project_datatable_browse.php?project=la_pp&datatable=vocabs
A GIS can include any form of archaeological data, which has a geospatial aspect or which can be associated with another geospatial entity. As such, contexts, loci, pails, excavated artefacts or biofacts or samples can all be included in a GIS. Ideally, the GIS used for an archaeological project would be integrated with its information system from the outset. One way of doing this is to use identifiers, which are universal across both systems. For example, the context entity with the label ‘1’ in a GIS is linked to the context ‘1’ in the linked information system. Another and perhaps more straightforward approach to linking GIS and all other archaeological project data is to use the GIS as the project datastore from the outset (Conolly & Lake 2006 pp. 51–60). This makes sense as a GIS is essentially a database, albeit one with a geospatial user interface.

The Priniatikos Pyrgos project employed the services of a GIS specialist from 2007 (Megarry 2014). Initially, ArcGIS (Law & Collins 2013), which is the industry leading platform, was used to spatially record and represent the site stratigraphy and material record. In 2010 the project abandoned the use of ArcGIS and instead adopted the Quantum GIS platform (Sherman 2008), principally because QGIS is Open-source Software and, therefore, free of any licensing expense.\textsuperscript{164} The primary job of both

\textsuperscript{164} Financial expense has now been listed on a number of occasions as a reason why a decision was made regarding the management of the data of the Priniatikos Pyrgos Project. Often, in circumstances where financial implications are weighed up against utility, finance ultimately proved the determining factor. This is a very common phenomenon in research archaeology, particularly in the wake of the global financial crisis of 2007-2008. As in the case of the switch...
iterations of the project’s GIS was to record and represent the positions of all project contexts as geospatial polygon data (Figure 103).

However, despite this early investment in both an Archaeological Information System and a GIS, the two were never integrated into a consolidated whole. So while the records in the FileMaker contexts table did contain geospatial plans, these representations were rendered as unstructured JPEG image files, which had been exported from the GIS and so the level of integration between the two datasets was minimal. It was only in the latter stages of the creation of the project’s RDF representation, that an attempt was made to integrate the GIS and Archaeological Information System into one unified system. The spatial data representing the project contexts was exported from Quantum GIS as a shapefile, a common vector file format used for geospatial data. Each context was represented as a polygon of points, with each point being made up of an X, Y, and Z coordinate relative to the WGS 84 coordinate system. Quantum GIS was used to access each of these points and to export them as polygons within a KML file. Lastly, the KML node associated with each context in this KML file was extracted using a Python script and this data was inserted into the Open Refine contexts project data (Figure 104).167

Figure 104: a screenshot showing the KML data for context 1 in Open Refine

from ArcGIS to Quantum GIS cited here, the day-to-day practice of the affected archaeological team is more often than not disadvantaged to some extent as a result.

This is not to say that efforts to integrate the two datasets were non-existent. For instance, at one stage data was exported from the FileMaker Archaeological Information System as CSV and then imported into the GIS in order to produce distribution maps for certain artifact types and periods. However, these experiments were never brought beyond the experimental stage.

The World Geodetic System (WGS), whose latest version is WGS 84, is one of the most commonly used coordinate systems in modern cartography (Slater & Malys 1998).

More recently, these KML data have been converted into the GeoJSON format as well.
Unfortunately, the locations of the project loci were never recorded in a GIS and so an alternative solution was needed to generate this information and by inference, all of the other archaeological entities (pails, artefacts, et cetera) associated with them. The solution found was to assign the polygon for each locus’s parent trench to the locus itself. While not ideal, in the sense that it was a less precise reading of the loci’s geospatial positioning, there was no alternative to recovering what was essentially absent data. Choosing to ignore the geospatial data for all loci and their associations would have excluded them from all future spatial analyses of the data and given that around half of the total material excavated from the site comes from these loci, this would have been a needless concession. Once again, pragmatism ruled the day.

As we will see in Chapter 6, having an ability to geospatially locate any archaeological object relative to any other archaeological object at the site, by its association with a context or locus and that context or locus’s KML data, became crucially important when it came to analysing the Priniatikos Pyrgos dataset. As was noted above, archaeologists tend to view the data that they study in terms of its spatial relationships to other data and in order for the Priniatikos Pyrgos to be able to deliver this type of analytical functionality, it was necessary to inject this geospatial data into its web of data.

**Processing image data**

The mapping of image data from the FileMaker database into the linkedarc.net system presented its own unique set of challenges. Priniatikos Pyrgos images came in the form of digital scans of plan drawings and digital photographs with the latter being by far the most numerous. Photographs were created digitally both on and off site at Priniatikos Pyrgos and it was the job of the archaeologist in charge of the particular photograph to label its filename following a particular convention depending on its subject matter. All photographs taken on site, which were known as field photos, were logged in the central site register and these identifiers were then used as the digital filenames. The second type of project photograph, the catalogue photograph, was given a filename based on the identifier of the catalogued object being captured (Figure 105).

---

168 This approximation of the relative magnitudes of the parts of the sites excavated using the SC method compared to those excavated using the LP method is based on their being 169,874 sherds excavated from Priniatikos Pyrgos contexts, while 174,990 sherds were excavated from loci. These figures were generated by sending a SPARQL query to the linkedarc.net endpoint.
In FileMaker, the images table includes contextual information describing the photograph’s subject and technical details such as the camera used and the direction faced, if it was taken in the field (Figure 106). It was possible to import all of this data, excepting the image itself, into Open Refine for cleaning, as is described above for the other tables. However, none of the FileMaker export formats allowed for the extraction
of the image table’s image data and given that images constituted a sizeable proportion\(^{169}\) of the Priniatikos Pyrgos dataset, this was a problem that required a solution.

All of the image files, which had been imported into FileMaker, were also stored in their original uncompressed state as files in a file system. These files were all uploaded to the Flickr image hosting service (Figure 107).\(^{170}\) Flickr is an extremely useful tool for the archaeologist looking to archive large amounts of image data.\(^{171}\) The Priniatikos Pyrgos Project had accumulated around 10GB of uncompressed image data and in order for these resources to become part of the project’s RDF dataset, they needed to be housed online and to be given URIs. Each image file uploaded to Flickr is given a permanent URI with which to access it. Flickr also provides access to all of its hosted material in a number of different resolutions. These variants can be accessed by appending a suffix to the original image URI (Flickr 2015).

---

\(^{169}\) In data size terms, images accounted for over 90% of the FileMaker database file.

\(^{170}\) [https://www.flickr.com](https://www.flickr.com)

\(^{171}\) At the time of writing (April 2015), Flickr granted each of its users 1TB worth of storage on its servers.
Flickr also provides a web API, which can be used to retrieve the original filenames and new URIs associated with each photograph in a gallery. This API was used to retrieve a list of all 9,057 Priniatikos Pyrgos photographs and this information was then merged in Open Refine with the images table data exported from FileMaker (Figure 108) and the result is shown in Figure 109.\textsuperscript{172}

\textbf{Outputting the project data as RDF using RDF Refine}

The final stage of the cleaning and mapping process for the Priniatikos Pyrgos dataset involved the creation of the RDF data, which was then used to populate the linkedarc.net triplestore. As discussed earlier in this chapter, the Priniatikos Pyrgos

\textsuperscript{172} Note that the use of Flickr as an external store for a large portion (in data size terms) of the Priniatikos Pyrgos dataset raises questions about the sustainability of such a strategy. Is it sufficient for an archaeologica l project to assume that a commercial company such as Flickr, which is susceptible to the vagaries and occasional turmoil of the market, will remain in existence for an amount of time that the project deems adequate for the function that it performs? And what might such a time period be?
dataset was mapped to a combination of data ontologies: CIDOC CRM, CRM-EH, the custom linkedarc.net and Priniatikos Pyrgos ontologies and other miscellaneous public ontologies (FOAF, SCHEMA, DCTERMS).

Both the CRM and the CRM-EH are complex models and the mapping of the multiple Open Refine projects, each corresponding to one of the original Priniatikos Pyrgos datatables, was challenging in its breadth. Luckily, the RDF Refine extension for Open Refine eases this burden with its intuitive user interface design (Figure 110). The user edits what is known as an RDF ‘skeleton’, which is essentially a mapping schema. This allows the data designer to decide how the fields in the Open Refine project are mapped to RDF triple patterns in the output. The first decision that needs to be made when constructing the RDF skeleton is to choose a URI address on which all the RDF outputs are based. In the example shown in Figure 110, the mapping uses a base URI of <http://vmubuntu/data/la_pp/>. This means that all of the RDF subjects created by this mapping will begin with <http://vmubuntu/data/la_pp/>. In practice, each separate Priniatikos Pyrgos datatable includes an extra string at the end of this base URI, which identifies its data type. For instance, a context with the name ‘1’ would be given the subject <http://vmubuntu/data/la_pp/context1>.

173 You may be wondering why the base URI reads <http://vmubuntu> and not <http://linkedarc.net>. This is because the RDF data outputted by the Open Refine Priniatikos Pyrgos projects is first deployed within a development linkedarc.net server configuration, which is run on a local virtual machine, whose base URI address is http://vmubuntu. All references to ‘vmubuntu’ are replaced with ‘linkedarc.net’ within the test linkedarc.net dataset whenever the development server replaces the live server. We discuss the test and deployment process at the end of this chapter.
The main body of the interface is divided into three vertical columns, which represent the subject-predicate-object structure of the RDF triple. The subject is entered into the left column as a root node of a tree of associated predicate and object pairs. RDF prefixes can be imported into any RDF Refine project skeleton. This hugely simplifies the entering of mappings and condenses the serialisations that are eventually outputted. Once added, these prefixes become available to the user allowing URIs to be entered in abbreviated form. For example, la:project can be entered in place of <http://vmubuntu/ontology/project>. RDF Refine’s prefix functionality further eases this process by prompting the user with a list of possible values through its autocomplete function. When dealing with the amount of classes and predicates contained within data models such as the CRM, this is a very useful feature, as it protects against the mistyping of entries. Incorrectly entered URIs at this stage of the mapping process is a cardinal error and one, which can be extremely difficult to track down after the event.

![Diagram](image-url)

**Figure 111:** the mapping of Open Refine table data to an RDF Turtle serialisation

Predicates are entered in the central column and object values in the final column. These object values could be constants, as in the case of the object value for the la:project predicate shown in the first entry of Figure 110 but they are more likely to be links to fields found within the Open Refine project. By making a project field the object of any triple combination, the RDF Refine extension creates a new triple for each row contained within the Open Refine project. As such, changes to the project’s data result in the creation of updated triples (Figure 111).
In its simplest form, an RDF mapping such as this would contain a single subject for each table entry. This is accomplished by adding a single root node to the skeleton. In this scenario, each outputted subject would be constructed from an entry’s row identifier. This subject would then be linked to a tree of predicate and object pairs that correlate to the table’s fields in a one-to-one fashion. However, due to the complexity of the class hierarchy needed to model the Priniatikos Pyrgos data, it was often necessary to create more than one subject in each Open Refine project’s RDF skeleton.

You will find in the appendices at the end of this text a detailed account of the particular mappings used for each of the Open Refine projects. While the basic method is repeated throughout and follows the overview provided above, each project contains its own particularities. While some of the mappings, for example, of the period data, are relatively straightforward, others, such as the context data mappings, are much more involved and create a more complex RDF graph output.

Serialising as RDF triples

Having mapped all of the source fields onto a combination of CRM, CRM-EH, linkedarc.net and Priniatikos Pyrgos ontological classes and predicates, the final step in the Priniatikos Pyrgos mapping process was to export the data as RDF. The RDF Refine extension allows you to export your data as either RDF Turtle or RDF/XML files. Each of the Priniatikos Pyrgos datatables created its own outputted RDF file and a sample of...
the contexts data serialised as RDF Turtle is shown in Figure 112. As has been mentioned already, the mappings defined in the RDF skeletons, do not for the most part produce a single RDF subject type as an output. For example, the images project creates RDF triples associated with subjects of the la_pp_ont:LA_E18_Image class. However, it also creates links between these la_pp_ont:LA_E18_Image entities and context, sub-context, find, locus and pail RDF subjects. As such, it makes more sense at this point to conceptualise the Priniatikos Pyrgos dataset as a cloud of interlinked resources (Figure 113). The previous mental image of the dataset as a series of hierarchical datatables is no longer an adequate representation of its form. And I would argue that this networked understanding of the output of the archaeological investigations undertaken at Priniatikos Pyrgos better reflects the way that archaeologists actually think about their data.

![Network visualization of Priniatikos Pyrgos dataset](image)

**Figure 113:** a visualisation of the network created by 100,000 of the 5,985,261 triples in the Priniatikos Pyrgos dataset

### Test and deployment servers

Once created, the Priniatikos Pyrgos RDF data was tested using a test server infrastructure. This test setup was built within a virtual machine environment, which runs on the VMware Fusion platform (Nieh & Leonard 2000). Hardware virtualisation has become an increasingly popular way of building server infrastructures (Crosby & Brown

---

574 This is a network visualization created using the Gephi tool. The 5,985,261 triple figure is correct as of 3 June 2015.
They allow for the division of a physical piece of hardware into multiple emulated pieces of hardware. From the perspective of the software, which runs on them, these virtual hardware stacks are almost indistinguishable from ‘real’ hardware. Hardware virtualisation delivers huge cost and energy savings when compared against traditional hardware setups. The approach has become a foundational strut in what is known as cloud computing (Armbrust et al. 2010). Hardware virtualisation is also the ideal platform on which to test server configurations. In the context of the linkedarc.net project, it allowed the system to be developed and tested using a single machine. This significantly reduced the costs and complexities involved in the development process.

The linkedarc.net virtual machine configuration is a mirror of the deployment server’s configuration. It runs Ubuntu ver12.04, a Linux operating system, which is highly popular among administrators of web servers. It employs an Apache 2 stack to deliver its web services. The Apache Jena Fuseki stack, which provides the RDF functionality for the linkedarc.net server infrastructure, sits on top of this configuration. Fuseki is shipped with a set of Fuseki scripting files, which can be used to import content into the Jena triplestore. The s-put script overwrites the contents of the triplestore with the specified RDF file while s-post adds the contents of the specified RDF file but does not overwrite the existing contents.

Once the RDF data has been thoroughly tested and validated within the test setup, it is promoted to the live deployment server, where it takes the place of the live dataset for the project on the linkedarc.net site.

---

175 It is very difficult to obtain figures about the share of operating system use within the server market. However in broad terms, based on a report released in 2012 by IDC (Nagel 2012), a market research firm, Linux and Unix-based systems accounted for just under 40% of server operating system revenues. It is difficult to extrapolate from this figure a corresponding unit figure but it is likely due to the lower cost of Linux and Unix licenses relative to their leading competitor, Windows, that the actual unit amount is higher than this revenue figure suggests.


177 Note that running these data update scripts when users are querying the Jena Fuseki triplestore may produce unexpected results.
On inferences

The answering of queries by drawing inferences on the subject dataset is an important capability of Archaeological Semantic Web data resources. We have already introduced the idea of inferencing in Chapter 3. Basically, inferential logic works on the principle that a query asking about a particular class will include all instances of subclasses of that class as well in its response. For example, if a dog is a subclass of an animal, then a dog is also an animal. Thus, if a query asks about animals, instances of the dog class should also be considered. In the context of the Archaeological Semantic Web, SPARQL is the ideal mechanism by which inferencing can be realised.

Given that the Priniatikos Pyrgos dataset is modelled upon an ontology that extends the CRM and the CRM-EH, which in itself extends the CRM, the ability to rely on inferences becomes vital when the dataset comes to be queried. For instance, the la_pp:LA_E18_Image is a subclass of the ecrm:E38_Image class. Without the support of inferences, were a user to ask the Priniatikos Pyrgos dataset for a list of its ecrm:E38_Image instances, it would return nothing in response, even though there are 9,057 instances of the la_pp:LA_E18_Image class contained within the dataset.

```
<http://linkedarc.net/data/la_pp/image2805-10778923523_25176479ef> a
  la_pp_ont:LA_E18_Image,
  ecrm:E77_Persistent_Item,
  ecrm:E36_Visual_Item,
  ecrm:E89_Propositional_Object,
  ecrm:E28_Conceptual_Object,
  ecrm:E1_CRM_Entity,
  ecrm:E71_Man-Made_Thing,
  ecrm:E73_Information_Object,
  ecrm:E72_Legal_Object,
  ecrm:E70_Thing,
  ecrm:E90_Symbolic_Object,
  owl:Thing,
  ecrm:E38_Image
```

Table 21: adding inferencing support to a Priniatikos Pyrgos image resource

The linkedarc.net Apache Jena Fuseki engine provides support for the answering of SPARQL queries using dynamic inference generation. However, this places a substantial processing load on the server and as a result this approach is unworkable in practice. There is, however, an alternative. It is possible to manually add the triples needed to encode the inference meaning. In the case of subjects of type la_pp:LA_E18_Image, this can be achieved by creating a triple that links the subject to the ecrm:E38_Image class. This is done because ecrm:E38_Image is the parent of la_pp:LA_E18_Image. Following
the same logic, the subject could be linked to ecrm:E36_Visual_Item, which is the parent of ecrm:E38_Image, and so on. This would create the set of triples shown in Table 21, which, when added to the triplestore, essentially delivers inferencing support.

**Digitally reassembling the Priniatikos Pyrgos team**

Through the course of six excavation and four study seasons, over one hundred people worked with the Priniatikos Pyrgos project. At its peak in 2010, the project hosted somewhere in the region of fifty individuals at any one time. This group was made up of undergraduate and postgraduate students, academic professionals and volunteers working both on and off the site in excavation and in the post-excavation process. While large and undoubtedly complex, this would not be unusual for archaeological field projects working in the eastern Mediterranean region (Davis 2007). Archaeology has changed in many ways over the years and this thesis has highlighted a number of these transformations but its dependence on large quantities of labour has diminished little. At Priniatikos Pyrgos, it was important to keep track of exactly who was responsible for each of the thousands of individual tasks that were undertaken during these ten years and it is the final broad data type that we need to consider in this review of how the Priniatikos Pyrgos data came to be modelled as RDF.

Like many archaeological field projects, Priniatikos Pyrgos employed the use of an initialling system as a means of associating an employee of the project with a particular task. Everything from the excavation of contexts in the field to the reading of pottery and cataloguing of individual pieces in post-excavation was accounted for by the inclusion of a set of initials on the relevant paper record. This information was translated into the digital domain as field data within the FileMaker database structure. For example, context records included the fields ‘Recorded by’, ‘Excavators’, ‘Sheet checked by’ and ‘Entered in DB by’ and pails, sub-contexts, catalogue pottery and objects tables all contained similar fields.

All of this agent identifier information was contained within the RDF created by the mapping of the other datatables. If you recall, many of the CRM events modelled by these mappings included links to instances of the crmeh:EHE0077_ProjectTeamMember class, representing the individuals involved. The objective of this next process was to establish a minimum set of these initials. Once all of the other project data had been imported as RDF into the linkedarc.net triplestore, the SPARQL query shown in Table
22 was sent to the linkedarc.net SPARQL endpoint in order to generate a list of the minimum set of individuals referenced by other RDF resources in the system.

```
PREFIX crmeh: <http://purl.org/crmeh#>
SELECT DISTINCT ?actor
WHERE {
  ?actor a crmeh:EHE0077_ProjectTeamMember .
}
```

Table 22: a SPARQL SELECT query used to generate a list of the Priniatikos Pyrgos team members

The data gleaned from this query was then imported into Open Refine to create the team members project. Most of the values used in these fields conformed to the rules of the initialling system but many did not. A common problem witnessed was the use of first names in place of initials and there were also the inevitable typing mistakes. Because of these problems, it was necessary to thoroughly clean the data. Then each of the initials needed to be associated with a first and a family name. Given the quantity of individuals involved, it was decided that the most efficient way of gathering this information would be to put out a call to all past members of the Priniatikos Pyrgos Project. A spreadsheet document was created and shared on Google Drive and the Priniatikos Pyrgos Project Facebook page was used to encourage past members to annotate the document with their details.\(^78\)

The initialling system is an example of a data type, which was originally designed to satisfy a particular need (that of keeping track of changes made to paper records) but with a small amount of effort can be reimagined with a new research focus. It is easily forgotten that archaeological finds are outcomes or products of human agency, both past and modern. They are revealed through the anthropocentric process of excavation. They are extracted with care, fastidiously analysed and in exceptional cases they can be presented to the public as part of a museum collection. The digital preservation of these social links offers the potential for considering the dataset as a sociological system. What were the nationalities of the people who worked on the excavation? How might the interpretations arrived at by these different nationalities have differed? Can these vagaries be detected in the digital record? How did these individuals interact with the data and by

\(^{78}\) As of 19 September 2015, 121 of 308 unique initials were associated with first and last names by this process. The Priniatikos Pyrgos Project Facebook group, which is accessible at https://www.facebook.com/groups/126842811810 was established in 2009 and it contains over 100 members, the majority of which have worked with the project.
inference with one another? Did individuals return to work at the site over a number of seasons? And so on. The asking of these sorts of questions falls outside of the traditional remit of core archaeological enquiry. However, in recent years archaeologists have shown the value to be gained by engaging with these alternatives perspectives (Caraher et al. 2014; Morgan & Eve 2012; Stephen & Morgan 2014).

**Conclusions**

I hope that this chapter has demonstrated just how complicated and involved the mapping of an archaeological dataset to an ontological model such as the CIDOC CRM can be. The sorts of issues encountered (processing of messy free-text data, the handling of multiple data types, the extraction and creation of structured vocabularies or the selection of semantic mappings) are not unique to this project (Carver 2013; Kansa 2014), although there are certainly characteristics of this particular data scenario, which did not help to ease the overall burden. As such, it is worth reflecting now on the process from a distance and commenting with the benefit of hindsight just how the workflow might have been improved upon, were it to be started again. Hopefully, these observations will also be relevant for other archaeological projects looking to achieve the same outcome.

Undoubtedly, the majority of the issues described in the chapter came about as a result of the form of the input data. In fact, it could be argued that a number of these problems can be traced even further back in the information chain to the drafting of the Priniatikos Pyrgos paper records themselves. The implications of using free-text fields have been repeatedly referred to in this chapter. Based on this experience, I would not advocate a blanket ban on the use of free-text fields in a database design; certainly, particular types of information such as descriptions would be difficult to accommodate without using free-text fields. Having said that, whether you choose to use these types of fields or not, is a decision which should be consciously taken and not one, which merely represents the route of least resistance.

Ideally, all data modelling decisions should be outlined in a document and the arguments for and against any particular design decision should be set out accordingly (Alhajj 1999; Denard, Hugh 2012 p. 66). This would help in the decision-making process, as it would force the data designer to consciously address the pros and cons of any one element. Documenting a database’s design also has the clear advantage of recording the design
process. This will enable future consumers and modifiers of the database to access this thought process, so as to better understand the database’s structure and content.

The subject of ontologies provided a central narrative feature within the chapter and this was no accident. Ontologies represent a fundamentally important aspect of any data mapping process. As such, they should be at the forefront of any Archaeological Information System design process. The Priniatikos Pyrgos Project dataset was not originally designed with the CRM and its CRM-EH models in mind. However, it was planned in general accordance with MoLAS’s interpretation of the Single Context excavation and recording method (Museum of London 1994), and it is on the SC method that the CRM-EH ontology is ultimately based. And this fortunate coincidence undoubtedly proved helpful during the mapping process. For instance, the information contained in the context sheets used by the project largely mirror the data structure centred on the crmeh:EHE0007_Context class. On the other hand, there were many other types of recording sheet employed at Priniatikos Pyrgos, such as the sub-context sheets, catalogue records and the sample sheets that did not have this same direct link to the CRM-EH structure. And then there was the additional and not insubstantial problem of accommodating Locus-Pail records, which were used to record one half of the site. This all culminated in a complicated mapping process, which would have been significantly attenuated had a data model, such as the CRM-EH, been used to structure the project’s database from the beginning.

A related problem was the use (or the lack of use) of vocabularies by the project during the paper records and early digital phase. We have already addressed the challenges of dealing with free-text data and the problems associated with controlled vocabulary mapping are tightly bound up with this discussion. As we have seen, vocabularies are a vital element in the proper workings of an Archaeological Semantic Web resource such as linkedarc.net. They open up a dataset to be intelligently queried by users who might initially know very little about its makeup. In this sense, they represent an entry point to a path of discovery. For example, a user might look upon the initially confusing mesh of information that is the Priniatikos Pyrgos linkedarc.net dataset and first ask whether it holds data this is of type ecrm:E39_Actor. They do so knowing that the dataset supports the CRM model and that ecrm:E39_Actor is the canonical way within that ontology of modelling people or groups. Having found a list of subjects of this type, they can then
explore the properties associated with these resources and from there proceed to construct more complex and useful questions for the dataset.

Vocabularies, therefore, should be the second (after data ontologies) consideration on the data designer’s mind during the planning stage of a new database. For example, if the planned database is to record contexts, the data associated with these entities should include a type field that is controlled; in other words, the person creating this record must select a type for the context record from a list of predetermined values. This approach is open to the charge of determinism and essentialism, but in certain circumstances the data designer may be required to choose utility over intellectual freedom. These controlled vocabulary lists should be collated by the data designer through a process of research and reflection on the aims and character of the dataset being created. Resources such as Seneschal and the Getty, both described in Chapter 4, should be the first port of call when collating these lists. These lists should then be compared against the vocabularies used by large cultural heritage digital content providers such as the British Museum to see whether there is general agreement between the two sets. This is to promote best practice. By observing how frequently particular vocabulary values appear in existing Archaeological Semantic Web resources, the data designer will be able to determine their level of familiarity within the archaeological community. And by deciding to use these resources, the designer will be able to co-opt this level of understanding into the new database resource.

Lastly, we need to consider the issues of data size and scalability in the context of Archaeological Semantic Web mapping operations. Firstly, on the point of scale, the Priniatikos Pyrgos dataset currently contains 5,985,261 RDF triples.\footnote{As of 3 June 2015.} This dataset is hosted on a server with 6GB of RAM, which can comfortably handle a dataset of this size. Further investigation is needed to ascertain the upper limits of Apache Jena’s carrying capacity but previous studies would suggest that this figure is significantly higher than 4,259,534 (AndyS 2008; W3C Wiki 2015).\footnote{This figure is correct for the Priniatikos Pyrgos RDF dataset as of 11 September 2015.} However, while triplestores may be capable of accommodating these larger datasets, scaling up an existing archaeological project is about more than simply triple counts. As has been shown in this chapter, mapping archaeological data from a paper or ‘closed’ digital form onto an RDF ontology
can be a time-consuming and demanding process; one requiring the application of significant amounts of domain knowledge and technical expertise. Were this investment to be required on an on-going basis (for example, after additional archaeological field seasons), it would be debatable whether the investment was warranted on the basis of a cost-benefit analysis. Thankfully, certain technical facilities exist to manage this update process (For instance, Open Refine includes an ability to record the data cleaning and mapping steps taken by a project. These can then be replayed on new source data) but ultimately, the best solution to this problem is to move these processes closer to the trowel’s edge. This would mean that controlled vocabulary and public ontology alignment be embedded into the creation of the archaeological data as close as possible to the source of its conceptual creation, as has been discussed earlier in this conclusion.

---

181 See Isaksen’s argument against the feasibility of the Archaeological Semantic Web on this basis (Isaksen 2011).
Chapter 6 – Mining the Archaeological Semantic Web

Asking questions of linkedarc.net and the British Museum

‘See now the power of truth; the same experiment which at first glance seemed to show one thing, when more carefully examined, assures us of the contrary.’ (Galilei [1638] 1954 p. 164)

Introduction

In Chapters 4 and 5, we considered the Archaeological Semantic Web as a concern of the data provider. Chapter 4 outlined the design of the linkedarc.net server and web app, and Chapter 5 explained how the data of the Priniatikos Pyrgos Project was turned into RDF data and imported into the linkedarc.net triplestore. Chapter 6 completes this narrative trajectory by introducing the archaeological data consumer into our study for the first time. The aim of this chapter is to consider how Archaeological Semantic Web resources might be interrogated as part of an archaeological research workflow. We begin by asking how the stratigraphy of the Priniatikos Pyrgos site might be accessed using SPARQL. We then present a visualisation of the results of these stratigraphic queries using the Gephi tool. In our second case study, we consider the loomweight assemblage of Priniatikos Pyrgos. In this investigation, we describe the PPTextiles web app, which presents the spatial distributions of the loomweight assemblage at the site. The Priniatikos Pyrgos loomweights are one of the more interesting categories of material culture found at the site, as they represent the physical remains of a single activity type found across time and space. Using the PPTextiles web app’s visualisation of the artefacts’ distributions, it is possible to conduct a diachronic case study of the textile production activities of the site across four thousand years.

The third case study considers data, which is not associated with the linkedarc.net project. As was explained in the review of existing Archaeological Semantic Web resources in Chapter 4, the British Museum has played a pioneering role in the delivery of RDF resources to the cultural heritage community. Its dataset is extensive and presents a highly comprehensive mapping of the CIDOC CRM. The content and semantic richness offered by the British Museum’s digital collections present enormous
potential to the archaeological data consumer, and in this case study we look at the museum’s extensive collection of almost 116,000 digital cuneiform inscription records.\textsuperscript{182}

**Returning to Priniatikos Pyrgos**

The previous chapter’s account has shown that publishing the data of any archaeological project to the Archaeological Semantic Web, regardless of the project’s size or the material culture that it is concerned with, is no trivial matter. The selection and creation of ontologies, the alignment of controlled vocabularies, the cleaning and preparation of messy data, the handling of different types of media and the management of complex data mappings all take time and demand the investment of significant resources. Given the scale of this effort, you might be forgiven for asking why an archaeological project, potentially operating under resource constraints, would be inclined to stretch those resources by pursuing this strategy. What does it have to gain in real terms as a result of this work?

We now look at the Priniatikos Pyrgos digital record from this more consciously critical perspective. Exactly, what is it that RDF publishing brings to the practice of archaeological research? What new insights can it accommodate? How, if at all, is an RDF Priniatikos Pyrgos dataset different from its FileMaker predecessor? We begin this account of the Archaeological Semantic Web as the tool of the researcher by considering how it is that archaeologists actually ask questions of digital resources in the first place.

**Asking questions of Priniatikos Pyrgos on linkedarc.net**

It was only in the latter stages of the Priniatikos Pyrgos Archaeological Semantic Web project that the significance of SPARQL as a research tool became apparent. At first, it appeared to offer a somewhat esoteric view onto the data and this complicated impression was exaggerated when compared against other apparently more intuitive interfaces. Why learn SPARQL’s peculiar set of syntax rules, when it was far easier to use the familiar interface of an application like the linkedarc.net web app to access the Priniatikos Pyrgos dataset? One way of looking at this question is to consider it through the prism of intellectual freedom. While the linkedarc.net search and browsing interfaces are designed to be as flexible to the needs of the research user as possible, in reality querying interfaces such as these will always end up guiding the user in some way.

\textsuperscript{182} This figure is correct as of 23 June 2015. All the amounts referenced in this chapter relating to the British Museum’s cuneiform digital collection are valid of this date.
Whether that guidance is explicit or not is largely irrelevant. The outcome is a path through the source material, which is predetermined by a system. There is nothing particularly new in this observation. Scholars have always operated within systems of learning (Machlup 2014 pt. 1). We access the information contained within the physical space of a library within a particular set of parameters (Jakobovits & Nahl-Jakobovits 1987). The arguments that we consume in journal articles and other academic publication media are presented formulaically (Russell et al. 2009). We are conditioned to a model that demotes or even excludes source data from our published narratives. Language itself is the great determiner of knowledge’s dissemination. To strive towards a working environment that is absent of rules is a logical nonsense. Being aware of these constraints is, however, a different matter and it is one that when employed diligently can help promote the health of any knowledge system (Bourdieu & Wacquant 1992; Maton 2003).

Having said that, we now live in an age that is technically capable of overcoming many of the limitations that once applied to knowledge transmission systems. The flow of data from one agent to the next was once artificially constrained because no better way existed at the time to accommodate it. One of these technological changes is manifested in the form of scale; datasets are simply a lot larger in today’s archaeology. As these data amounts increase, so too does the range of question types that can be asked of this data. Another innovation has come about in the ways that information can be visualised and we will consider a number of these opportunities in the coming chapter. The interface between the human and a dataset is another crucial component of this system and modern technology has opened up exciting and perhaps even fundamentally revolutionary approaches in this space too (Brey 2005; Burdick & Willis 2011). There are now more choices available to the researcher and while interfaces like the linkedarc.net web app will satisfy certain types of enquiry and certain modes of work, others such as SPARQL come into their own in other research scenarios.

The following is an example of how one question can be posed of a linkedarc.net project in a number of different ways. The objective is to request from the linkedarc.net server a list of all of the Priniatikos Pyrgos contexts containing human bones. This is a fairly

---

183 Consider the case of the Google Books project, which has allowed scholars to ask questions of scales of literary information that were previously impossible (Fischer 2013).
typical type of archaeological question. To start, you could go to the linkedarc.net web app search page and enter ‘human bone’ into the search text field and click Search. This would present a list of results similar to those shown in Figure 114. This list can be browsed and individual records explored further by clicking on their hyperlinks. For example, you might wish to learn more about the Context 634 record. In the record displayed (Figure 115) you would discover that Context 634 was part of an ossuary known as ‘Grave 1’. You would also learn that a number of interesting finds were uncovered in the context, including 09-5803, a spectacular funnel necked glass flask with blown ribbing on its neck. You might then decide to search for other blown glass pieces found at the site from within the catalogued object display page. This would allow you to access the 366 artefacts in the dataset that match this description.
Now let us consider how the same question might be approached using the SPARQL method. The SPARQL query shown in Table 23 asks the linkedarc.net server for all of the context or locus entities, whose excavated contents included human remains. There is no hiding the fact that SPARQL can be a difficult paradigm to master. At first glance, it appears esoteric with the logic expressed by each constituent statement difficult to disentangle from the whole. With time and effort, however, SPARQL becomes a powerful tool in the hand of the archaeological researcher. It opens up datasets in ways that traditional digital search interfaces simply cannot, leading to new insights and new interpretations.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
</tr>
</thead>
</table>

Table 23: a SPARQL SELECT query asking for all contexts and loci containing human bones

Our first query is complicated, principally because the Priniatikos Pyrgos data structure, which it addresses, is modelled using the multi-level CIDOC CRM class hierarchy. It begins to make a little more sense, however, once you remind yourself that each of the SPARQL criteria follows the tripartite triple form. For instance, line 8 asks the SPARQL endpoint to return all subjects that are contexts (crmeh:EHE0007_Context) or loci (la_pp_ont:LA_E12_Locus). Lines 9-11 remove any subjects, which are not associated with human bone finds, by referencing the linkedarc.net vocabulary entry for human bone, la_vocabs:findtype-humanbone. Finally, lines 12-16 ensure that these find assemblages have contents and are not empty. Running this query returns a total of 18 subjects, of which 17 are contexts and only one is a locus (Figure 116). To further investigate any one of these subjects, you can click on their hyperlinks and from there you will be able to view all of the data associated with them.
Using SPARQL, we can choose to narrow down our search results by supplementing our original query with additional criteria. For example, we might insist that the context and locus entities returned not only have human remains associated with them but also carbon remains as is shown in the amended query reproduced in Table 24.

```
PREFIX ecrm: <http://erlangen-crm.org/current/>
PREFIX crmeh: <http://purl.org/crmeh#>
PREFIX la_pp_ont: <http://linkedarc.net/ontology/la_pp/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX la_vocabs: <http://linkedarc.net/vocabs/>

SELECT DISTINCT ?context WHERE {
  {?context a crmeh:EHE0007_Context} UNION {?context a la_pp_ont:LA_E12_Locus} .
  ?dimension ecrm:P91_has_unit 'bags' .
  ?dimension ecrm:P90_has_value ?value
  FILTER(xsd:float(?value) > 0)
} .
```
Table 24: A SPARQL SELECT query asking for contexts/loci with human bone and carbon remains

Now when we execute the query, the server responds with only 2 subjects, Contexts 3 and 786. This example has demonstrated how a relatively large amount of amount of data, which would have taken many hours to sort through by hand (for instance, using the linkedarc.net web app’s other search interfaces), has been whittled down very quickly from a total context/locus pool of 1,129 entities to just 2 entities that we are interested in. The point here is not that this end result would be impossible to achieve using an alternative interface – for it quite evidently could be found using a FileMaker Pro or other data management search mechanism – but that SPARQL provides a flexibility to navigate the often highly complex RDF graph networks built to model archaeological content. It is possible to imagine that any meaningful query that can be proposed by a user can be rendered using SPARQL. And this cannot be said for most dataset query interfaces, particularly those that contain data of multiple types. The second important point to make about SPARQL is that these queries are designed to interrogate Linked Data. All of the work that went into the construction of these links, which was described in Chapter 5, is now realised using a research mechanism that foregrounds the networked nature of these datasets.

We will now look at some specific examples of how SPARQL can be used to interrogate the Priniatikos Pyrgos dataset on linkedarc.net. Each case study will set out the research question broached, the context of that question, the methodologies employed, challenges met and the ultimate outcomes of the exercise.

\[ \text{?contextFind} \text{P45_consists_of la_vocabs:findtype-carbon .} \]
\[ \text{?contextFind} \text{P39_measured ?contextFindMeasurementEvent} \]
\[ \text{?contextFindMeasurementEvent} \text{P40_observed_dimension ?dimension} \]
\[ \text{?dimension} \text{P91_has_unit 'bags' .} \]
\[ \text{?dimension} \text{P90_has_value ?value} \]
\[ \text{FILTER(xsd:float(?value) > 0)} \]

---

184 In the relational database model, these multiple types would typically be stored in different tables within a database, rendering the construction of queries that target this multifaceted data much more complex.
Visualising the stratigraphy of Priniatikos Pyrgos

linkedarc.net resource

Detailed notes on the Priniatikos Pyrgos stratigraphy visualisation case study

http://linkedarc.net/datamining/gephi-pp-stratigraphy

Context

The stratigraphic method and the representations used to aid this particular analytical approach have a long and established tradition within archaeology (Trigger 1996 p. 127). In fact, so intertwined are the two entities that it would be difficult to conceive of the practice of modern archaeology in the absence of stratigraphy. The idea that material can be given a relative date by first identifying its context and then determining the relationship of that context to other contexts is fundamental to how archaeologists interpret excavated material. This view is predicated on the idea that contexts, which overlay other contexts, are younger than these lower layers. The law of superposition was first employed by geologists in the late 18th century. James Hutton was the first to use the axiom as a mechanism by which to explain and date the formation of layers of geological rock (Harris 1989 pp. 4–5). Soon after Hutton’s breakthrough, early archaeologists came to appreciate the potential of the law as an observational device within their own field (Daniel 1975 p. 25).

In the early years of archaeology, stratigraphy tended to be viewed from the horizontal perspective championed by the geologists. As such, the pioneering archaeologists of the late 19th and the first half of the 20th centuries placed a premium on their ability to excavate trenches not only as a means of accessing the artifactual material contained within them but also as a way of creating a series of baulks, which could then be read stratigraphically. We see this style of excavation in some of the earliest investigations that we now recognise to be archaeological in the modern sense of the word. General Augustus Pitt Rivers, for instance, adopted a system, which ensured that baulks be left as part of the excavation of the many earthworks, which he investigated at Cranborne Chase (Pitt-Rivers & Garson 1887; Trigger 1996 p. 293). He would then use the information exposed by these baulks when constructing his interpretations of the site.

This horizontal perspective reached its apogee in the mid-20th century within the practices of the energetic British archaeologist, Mortimer Wheeler. For Wheeler,
stratigraphy was the key analytical tool of the field archaeologist and he was insistent that all archaeologists should be masters of the skill as a baseline requirement (Wheeler 1954 p. 56). Wheeler is remembered for many reasons but it is hard to consider his contribution to the archaeological method without bringing to mind the so-called ‘Wheeler box’ trench layout system, which he innovated during his career (Wheeler 1955 p. 109). This system saw a site being divided up into a grid of 4x4m trenches, each separated by a 1m wide baulk. Each trench when excavated created four sections, which alongside all of the other sections exposed by the excavation of the other trenches constituted the primary source with which the site’s stratigraphy was elucidated. This ultimately provided the primary basis for the writing of the site’s central narrative.

While the Wheeler method of accessing stratigraphic material is still used to some extent in archaeological investigations today, it now tends to be either superseded or at least accompanied by a complimentary vertical viewpoint, which was first introduced by figures such as Philip Barker during the 1970s as part of the establishment of the Single Context excavation and recording method (Barker 1982). Around the same time, Edward Harris proposed a new diagrammatic approach to the representation of stratigraphic relationships (Harris 1979), which came to be known as the Harris Matrix method. This visualisation is now used extensively both within academic and commercial archaeological practice, particularly those that follow the SC method.

As we know from the accounts contained within Chapter 5, the site of Priniatikos Pyrgos was excavated using a combination of the SC and the Locus-Pail methods. As a result of the site’s re-occupation over a series of different periods and its exposure to significant levels of cultural and natural erosion during this accumulated time period, the reading of the stratigraphy of large sections of the site has proved immensely challenging to the site’s archaeologists. Throughout the five years of the site’s excavation, the project’s field archaeologists recorded the stratigraphic changes, which they encountered, on the paper context sheets and in their locus notebooks. This information was then digitised, firstly, as a part of a FileMaker database and, subsequently, in the RDF format as a component of a linkedarc.net project on linkedarc.net.

As already mentioned, the first objective of most archaeological stratigraphic analyses is to create a Harris Matrix. The creation of these interpretive aids for a section of a site
and ultimately to represent a site’s entire stratigraphic network constitutes one of the more demanding interpretive exercises that the archaeologist is asked to perform. Having said that, the interpretive rewards of doing this can be considerable. Harris Matrices can serve to guide and inform any further excavation at the site and they also provide a chronological map on which the site’s narrative can be constructed. To complicate matters, it is possible (and indeed highly probable) that contexts will be missed during excavation and only subsequent analysis of the context’s finds in conjunction with what is hopefully a thorough record of its excavation can rectify these mistakes. In practice, most stratigraphic information will go through a process of refinement and rationalisation throughout a project’s history. This iterative process was most certainly in evidence at the Priniatikos Pyrgos project where stratigraphic interpretations first made in the field were amended during the post-excavation process. This to and fro meant that the digital representation of this information be regularly updated within the linkedarc.net dataset.

Research question

This case study considers how Archaeological Semantic Web methods might be employed to highlight concentrations of material found within the complex stratigraphic record of Priniatikos Pyrgos’s Trench 2. The excavation of Trench 2 unearthed just over 400 separate archaeological events or contexts. Some of these units contained small amounts of artifactural material and/or biofacts,\(^\text{185}\) while others were associated with much larger quantities. An ability to quickly and intelligently extract relevant details from the vast amounts of data created by excavation, survey and material study is fundamental to the archaeological working practice. In this case study, we employ visualisation techniques using the Gephi platform in tandem with a SPARQL-based data-mining workflow to pursue this objective.

Methodology

Gephi is a free, open-source desktop visualisation tool written in Java and compatible with Mac OS X, Windows and Linux operating systems (Bastian et al. 2009). It was developed to fill an apparent tool gap in the visualisation of graph-based networks\(^\text{186}\) and archaeologists were quick to see its potential in the context of visualising various types of archaeological information, particularly those that are naturally inclined towards being

\(^{185}\) Biofacts are organic materials, which have archaeological significance.

\(^{186}\) Gephi allows for the visualization of tree hierarchies as well but structurally it addresses all of its input data as if it is polyhierarchical.
conceptualised as networks (Brughmans 2013; Graham & Weingart 2015; Larson 2013). Data is imported into Gephi using a variety of methods. Once in the system, the data is stored in either a nodes table or an edges table. The nodes table contains a list of entries, which represent the nodes within a network graph and the edges table entries define how the nodes relate to one another. As such, the data design is relatively simple to understand. The Gephi core system provides a certain baseline level of functionality but in order to fully maximise the package’s potential, it is best to utilise the sizeable array of plugins, which have been developed by a community of developers (Gephi Consortium 2012).

As long as you have a dataset, which contains related entries, Gephi can be used to visualise it as a graph. In that context, data structured as a relational database is an obvious candidate. In the same vein, RDF’s inherently networked structure provides another ideal data representation that can be visualised using Gephi. RDF data can be imported directly into Gephi using the Semantic Web Import plugin, which was developed by Demairy.\textsuperscript{187} This plugin sends a SPARQL query to a SPARQL endpoint, waits for the response and then populates the nodes and edges tables with data, which represent the RDF triples contained within the response.

Following the CRM-EH’s guidelines (May 2012), the crmeh:EHE1001_ContextEvent class is used to encode the stratigraphic relationships between two archaeological contexts. A number of different predicates can be used to encode the various possible stratigraphic relationships between crmeh:EHE1001_ContextEvent instances. The linkedarc.net implementation used just three of these predicates (ecrm:P114_is_equal_in_time_to, ecrm:P120i_occurs_after, ecrm:P120_occurs_before) to model this stratigraphic meaning for the Priniatikos Pyrgos data. ecrm:P114_is_equal_in_time_to is used to link two crmeh:EHE1001_ContextEvent

\begin{itemize}
\item \textsuperscript{187}This account will not include the specific procedural details involved in importing RDF data into Gephi by making SPARQL calls to an endpoint using the Semantic Web Import Plugin. A well documented guide on this subject already exists (Hirst 2012) and can be followed to access linkedarc.net RDF content. There is, however, one exception that needs to be taken into account when dealing with linkedarc.net data. The Semantic Web Import Gephi plugin is hardcoded to send HTTP POST requests to the specified SPARQL endpoint. The linkedarc.net server currently does not accept HTTP POST requests to its SPARQL endpoint. In order to overcome this incompatibility, it was necessary to include the entire SPARQL query in the plugin’s encoded URL parameter. An example request is as follows, http://linkedarc.net/sparql?query=SELECT+*%0D%0AWHERE+%7B%0D%0A%3Fs+%3Fp+%3Fo%0D%0A%7D+LIMIT+10&samplequery=default&output=xml
\end{itemize}
entities, which are deemed to be contemporaneous. ecrm:P120i_occurs_after and ecrm:P120_occurs_before link a crmeh:EHE1001_ContextEvent entity with another that is either earlier or later than it. Finally, crmeh:EHE1001_ContextEvent entities are linked to their related crmeh:EHE0007_Context entity via the ecrm:P7i_witnessed predicate. Using this ontological guide, it is possible to construct the SPARQL query shown in Table 25. This first asks the linkedarc.net server for the list of all contexts in Trench 2. It then uses the ecrm:P120i_occurs_after predicate to access their stratigraphic order.

```sparql
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX gephi: <http://gephi.org/>
PREFIX la_ont: <http://linkedarc.net/ontology/>
PREFIX la_pp_ont: <http://linkedarc.net/ontology/la_pp/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX ecrm: <http://erlangen-crm.org/current/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX crmeh: <http://purl.org/crmeh#>

CONSTRUCT { 
?context gephi:label ?cocontextName .
?contextBelow gephi:label ?cocontextBelowName .
}
WHERE { 
SELECT ?context ?cocontextName ?cocontextBelow ?cocontextBelowName { 
?trench a crmeh:EHE0088_SiteSubDivisionDepiction .
?trench ecrm:P87_is_identified_by "2" .
?context rdfs:name ?cocontextName .
?contextBelow rdfs:name ?cocontextBelowName .
?contextFind ecrm:P16i_was_used_for ?contextFindUseEvent .
?contextFindUseEvent ecrm:P4_has_time-span ?period .
?period ecrm:P87_is_identified_by ?periodName .
?period ecrm:P80_end_is_identified_by ?periodEnd .
} GROUP BY ?context ?cocontextName ?cocontextBelow ?cocontextBelowName
```

Table 25: retrieving the stratigraphic sequence of Priniatikos Pyrgos Trench 2 using SPARQL
Like the previous example, this SPARQL query appears daunting at first glance. However, once broken down into its constituent parts, it begins to make more sense and is not overly complicated. The SPARQL query is constructed of two main sections. The first employs a CONSTRUCT statement (lines 10-15), which is new to this thesis, and a SELECT statement (lines 17-34), which should now be largely familiar.

The list of criteria in the SELECT statement begins on line 19 by asking the SPARQL endpoint for all of its crmeh:EHE0007_Context entities. Lines 20-22 refines this list down to those contexts, which are contained within Trench 2. Line 23 retrieves the crmeh:EHE1001_ContextEvent entity associated with each of these contexts and line 24 finds the crmeh:EHE1001_ContextEvent entities, which occur before each of these entities. Line 25 links these related crmeh:EHE1001_ContextEvent entities back to their parent crmeh:EHE0007_Context entities. Lines 26-27 retrieve the names for the two stratigraphically related contexts. Lines 28-33 get a list of all of the finds associated with each context. It does this in order to retrieve the periods for each of these finds. The first line of the SELECT statement (lines 17-18) defines which variables are to be returned by the statement. Most of these are self-explanatory except for the last variable, which employs the use of the SPARQL MAX function. This finds the highest or latest date from the list of find period end dates retrieved for each of the contexts. This essentially gets the date for each of the contexts, as following basic archaeological context dating theory, the latest find date equates to the context’s date.

The CONSTRUCT statement takes the results of the SELECT statement and creates a set of new triples with this information. These triples are used to tell the Gephi tool how it should populate its node and edges tables. Lines 11-12 create a set of triples that relate the ?context and ?contextBelow variables to their labels. It employs the use of the <http://gephi.org/label> predicate, which tells the Semantic Web Import plugin to fill the ‘label’ column in the Gephi node table with the context names. Line 13 associates the context node with the latest period date, calculated using the SPARQL MAX statement on line 18. It uses a predicate called gephip:periodEnd to do this. This predicate has been created by the query and will appear as an entry within the Gephi nodes table within a column called ‘periodEnd’. This results in the creation of what is known as a Gephi attribute for the particular context entry. We will see in a moment how attributes play an important role in the Gephi system by allowing stylistic changes to be applied to graph
elements as a function of their value. Finally, line 14 creates a link between the context and contextBelow nodes within the Gephi edges table using the la_pp_ont:P13_matrix_isover predicate. Running this query using the Semantic Web Import plugin produces a Gephi data structure shown in Figure 117 and Figure 118.

![Figure 117: the resultant Gephi nodes datatable after running a linkedarc.net SPARQL query using the Semantic Web Import plugin](image)

![Figure 118: the corresponding Gephi edges datatable](image)

With the data collated in Gephi, it can now be visualised. The form that this visualisation takes is determined by the Gephi layout, of which there are a number. Different layouts produce different results and deciding on a layout depends on the research objectives. In this case and following the general approach to archaeological stratigraphic representations, we want to present a view of a network, which is displayed as a hierarchical up-turned tree. In this tree the latest root contexts should be positioned near the top of the structure. Older contexts should then branch out from these roots and
grow downwards. However, the only layout available, which ostensibly might deliver such a visualisation type, the ‘DAG Layout’, was unable to handle the complexity of the network represented by the Priniatikos Pyrgos Trench 2 dataset. As such, it was necessary to consider alternative and non-traditional layout schemas. In the end, the ‘Yifan Hu Proportional’ layout provided the next best alternative. This layout employs a force-directed graph algorithm whose goals are to minimise crossing edges, to produce edges of the same length, and to evenly distribute nodes across the graph canvas (Fruchterman & Reingold 1991; Tollis et al. 1998).

The results of applying the ‘Yifan Hu Proportional’ graph layout algorithm are shown in Figure 119. In this visualisation, the most networked nodes (i.e. those nodes with the most edge links) appear near the centre of the graph, while those with least links are
pushed to the periphery. In the context of the Priniatikos Pyrgos Trench 2 stratigraphic data, by dint of the fact that the latest contexts are most likely to be those with the most links to other contexts, these are positioned towards the centre of the graph, while the earlier contexts fan out towards the margins.

The data as represented by the ‘Yifan Hu Proportional’ layout produces an informative if somewhat unorthodox stratigraphic network, quite unlike the Harris Matrix’s upside-down tree form. The contexts found in a Harris Matrix are also often grouped together by period and this adds valuable analytical information to the visualisation (Figure 120). This type of information can be added to a ‘Yifan Hu Proportional’ type layout in the form of isolines. These link contexts of the same period and Figure 121 shows the results of one such attempt. The visualisation is moderately successful at highlighting the most prevalent periods in Trench 2. Contexts of Byzantine date dominate the central core of the graph, while large swathes of the periphery are given over to Late Minoan deposits.

The earliest periods of occupation at the site are clearly shown to be Early Minoan and secure contexts of this date are reasonably numerous. In general, the model, which places the latest contexts in the centre and the earliest contexts at the periphery, would appear to be supported by the addition of period isolines. The visualisation also highlights problems inherent within the data. An obvious issue is the large groupings of contexts, which are identified as being of unknown date. These darker areas are spread throughout...
the network and clearly indicate that the stratigraphic record for the trench is far from complete.

Figure 121: applying bands of colour, categorised by period, to the graph shown in Figure 119

Figure 122: a Gephi generated network graph describing the stratigraphic relationships between the Priniatikos Pyrgos Trench 2 contexts (colour-coded by period)
Figure 122 presents a slightly different view of the data. As an alternative to the period isoline approach seen in Figure 121, the nodes in this case are colour coded by period. While this approach does have certain advantages (it is arguably more aesthetically pleasing), the communications capability of this graph is limited as a result of its complexity. It simply contains too many nodes and edges to be easily legible using the print medium. One possible solution to this would be to create a series of graphs, each emphasising one particular period in turn, as is shown in Figure 123 for the Early Minoan deposits.

![Graph showing Trench 2 stratigraphic network with Early Minoan contexts brought to the fore.](image)

Another alternative would be to zoom in on portions of the graph as is shown in Figure 124. This has the advantage of being more legible but suffers from the fact that it is a partial view.
Thus far we have created a series of visualisations of the Trench 2 stratigraphic relationships, which is generally comparable to the results achieved using the Harris Matrix method. The graphs shown in Figure 119, Figure 121, Figure 122, Figure 123 and Figure 124 all deliver the same fundamental information as a Harris Matrix, albeit they do so using different representational conventions. The processes involved are, however, substantively different. The Gephi + Archaeological Semantic Web method is predicated on the idea that Archaeological Semantic Web resources can be re-used. The fact that data is exposed to the digital community in a way that allows it be consumed by a wide variety of different researchers, with varied questions in mind and tools at their disposal, is fundamentally different to the traditional pen and paper approach used in the creation of a Harris Matrix diagram. The point is not that one method is more valuable as a research tool than the other. The traditional approach to stratigraphic sequence creation is clearly of value to the archaeological process and using the Gephi + Archaeological Semantic Web approach does not remove the effort demanded in understanding the complex stratigraphic relationships of a site. This effort is interpretive and, ultimately, human interpretation must be where the primary investment of all archaeological practice is focussed (Redman 1987 pp. 249–251).

Having said that, technology, if employed intelligently, can help reinforce this reflexive approach to the archaeological process (Berggren & Hodder 2003 p. 425). If and when Archaeological Semantic Web practice is introduced as a research tool into the working archaeologist’s praxis, its flexibility has the potential of greatly opening up the range of questions that can be asked of aspects of an archaeological dataset. For instance, it could be argued that the Harris Matrix form can be misleading in certain respects as it presents
each context as a uniform entity positioned within a chronological sequence. In other words, all Harris Matrix contexts are created and presented equally. While this delivers a particular interpretive view of the data, it fails to take into account the possibility that individual contexts might be of higher relative importance when compared to other contexts based on a certain additional set of criteria. For example, it might be the case that the excavation of a particular Iron Age context yielded just a single sherd of diagnostic pottery, while a vast assemblage of Late Minoan ceramic was found in another. This is a meaningful distinction, which Harris Matrices in their orthodox representation tend to exclude. It is a reasonably trivial task to take the same basic workflow established to create the Gephi visualisations described above and to modify it slightly to inject this additional information. In this instance, the original SPARQL query could be amended as is shown in Table 26.

```sql
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX gephi: <http://gephi.org/>
PREFIX la_ont: <http://linkedarc.net/ontology/>
PREFIX la_pp_ont: <http://linkedarc.net/ontology/la_pp/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX ecrm: <http://erlangen-crm.org/current/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX crmeh: <http://purl.org/crmeh#>

CONSTRUCT {
  ?context gephi:label ?contextName .
  ?contextBelow gephi:label ?contextBelowName .
}
WHERE {
  SELECT ?context ?contextName ?contextBelow ?contextBelowName
  (MAX(?periodEnd) AS ?periodEndMax) (SUM(?nSherdCount) AS ?totalSherdCount) {
    ?trench a crmeh:EHE0088_SiteSubDivisionDepiction .
    ?trench ecrm:P87_is_identified_by "2" .
    ?context rdfs:name ?contextName .
    ?contextBelow rdfs:name ?contextBelowName .
    ?contextFind ecrm:P16i_was_used_for ?contextFindUseEvent .
  }
}
```
The linkedarc.net server is now asked not only to furnish the Trench 2 contexts and their periods but also to include the sherd counts for each of these deposits. This new information can then be used to create the visualisation shown in close-up in Figure 125.

The relative importance of each context in respect to its ceramic assemblage is now included in the visualisation as the size of the context's node. This makes this added and
important information clearly visible to the reader, which allows it to be absorbed and interpreted quickly.

Unsurprisingly, Context 1, as the topsoil deposit, is the largest in ceramic sherd count terms. A general pattern emerges, where deposit sizes reduce with distance from Context 1. This is, however, contradicted by contexts 26, 505 and 509, which have larger assemblages than their immediate neighbours. Therefore, following a central axiom of data analytics, it is with these anomalies that further investigation should begin (Han et al. 2011 p. 543).

**Research findings**

This case study produced a number of different visual outputs. Some of these build upon the conclusions of earlier visualisations, while others reflect the application of different approaches. Figure 123 showed that certain stratigraphic sequences within the trench are dominated by EM deposits (for example, those beginning with C117, C824 and C950), while others are completely devoid of EM material. Figure 125, on the other hand, augmented the trench’s stratigraphic record with information about the individual context’s ceramic weight. This served to highlight important contexts, such as C26 and C505, which stand out from their neighbours by the relative size of their associated material assemblages.

As an archaeologist who has excavated and studied the material at Priniatikos Pyrgos at first hand, I found that the most interesting outcome of the Trench 2 stratigraphy case study was the flexibility that the approach offered the researcher. Using the SPARQL/Gephi integrated workflow, I was able to quickly and easily update the visualisations, as and when the direction of my questioning evolved. While my original intention was to construct visual outputs that were in keeping with the established Harris Matrix visual convention, my outputs soon morphed into forms that diverged from this prototype. And by the end, I was employing quite different visual approaches to the representation of aspects of the Trench 2 stratigraphic record.

In many ways, the significance of the experiments conducted as part of this case study is not so much to be found in their final visual outputs but in the process of their creation. With that thought in mind – i.e. in foregrounding the importance of the medium over the message – it is not hard to imagine how a visual representational system such as is
described above might be expanded upon to develop this interpretive potential. For example, the Yifan Hu network graph form could be used as an initial user interface view onto a context-rich archaeological digital dataset. Users of the interface would navigate the data not via a tabular interface, as is employed in the linkedarc.net design, but instead using the more interpretively helpful Yifan Hu design framework. Because the linkedarc.net dataset can be accessed by the outside world via a SPARQL interface, shifts of fundamental research design such as this are made entirely possible, and that is a strong endorsement of publication as Linked Open Data.

The loomweights of Priniatikos Pyrgos

linkedarc.net resource

Detailed notes on how the PPTextiles web app was created
http://linkedarc.net/datamining/pptextiles

Loomweights are a category of material culture that is often found during archaeological excavations conducted within the eastern Mediterranean zone. For the most part, however, they are relegated to a brief comment in a publication’s ‘miscellaneous finds’ annex. This is regrettable as when considered alongside complimentary source materials such as the representative arts, spindle whorls and architectural installations, they can be used to provide valuable insights into a rich facet of ancient social and economic life, which is often overlooked: textile production (Wild 2007 p. 1). The study of loomweights also has many similarities with the practice of ancient ceramics analysis, which is by far the most common form of archaeological specialisation, particularly in the Eastern Mediterranean region (Dickinson 1994 p. 101). As with loomweight analyses, traditional ceramics studies is dominated by the language and practices of processual ways of thinking and working, and a consideration of the dividends forthcoming from a marriage between Archaeological Semantic Web approaches and ancient ceramics studies would no doubt warrant a thesis of its own. The following loomweight case study can be viewed as an exemplar of the more flexible epistemological process delivered by the exploitation of Archaeological Semantic Web resources, and many of the conclusions drawn could also be applied to ceramic material assemblages.
Chapter 6 - Data mining the Archaeological Semantic Web

Research question
As we know from the account of the site included in Chapter 5, Priniatikos Pyrgos has revealed evidence of occupation across a wide range of chronological periods. Loomweight artefacts have occurred in a large proportion of these deposits and, as such, they present an ideal material type with which to conduct a case study into changing patterns of the use of space at the site over a large expanse of time. This case study introduces the PPTextiles tool, a web-GIS interface built upon an Archaeological Semantic Web data-access layer. Our case study considers what this geospatial view onto the Priniatikos Pyrgos loomweight assemblage tells us about the site’s history of use.

Context
In the eastern Mediterranean, loomweights and spindle whorls are the most commonly excavated forms of material culture associated with past textile production and consumption behaviour. The dominant climate of the eastern Mediterranean, being generally cold and wet in the winter and dry and hot in the summer months, is not very conducive to the survival of organic materials within a soil matrix. As such, allowing for a tiny number of exceptions, the textile products themselves almost never form part of the material record recovered through excavation in this area (Burke 2010 p. 2; Gillis & Nosch 2007 p. vii).

And yet despite this lacuna in the sources, textile production and distribution would have constituted one of the major domestic, economic and social spheres of activity in the ancient world (Barber 2007; Jorgensen 2007 p. 7; Waetzoldt 1972). For instance, it is highly likely that the powerful economies of Bronze Age Crete and Iron Age Phrygia.

---

188 http://pptextiles.linkedarc.net
189 It is somewhat disingenuous to speak of a single climate type when addressing the Mediterranean region. If anything, the region is typified by a dissimilarity of regional character (Walsh 2013 chap. 1) and this leads to a range of climate possibilities. For the purposes of this account, we are referring, specifically, to the lowland, coastal and island zones of the region.
190 Çatalhöyük, an Anatolian site, has some of the earliest evidence of fabric impressions, dating to c. 6,000 BCE (Barber 1991 p. 127). On the Greek mainland and in the Greek islands there is an almost complete absence of fabric remains associated with Bronze Age and Iron Age contexts. For the most part we are left to derive what interpretations we can from fabric impressions made onto other objects such as ceramics, as with the plain-weave example found on a Middle Neolithic sherd from Sitagroi in Thrace (Renfrew 1972 p. 351). It is only by the 5th century BCE that we encounter physical textile remains within the archaeological record of the eastern Mediterranean (Moulherat & Spanidaki 2007).
191 The production of textiles was a vital economic cog in the worlds of both the Minoan and Mycenaean inhabitants of Crete (Militello 2007).
in Asia Minor were built in no small measure upon the proceeds of trade in textiles. For
the most part, we are left as archaeologists and students of the ancient social and
economic histories of southern Europe to consider the categories of evidence that speak
to us indirectly of these ancient textiles. One such source is the representative arts, which
in certain instances provide a valuable insight into the appearance of the finished fabrics
themselves. There are also examples depicting textile work areas and textile production
equipment. Figurative decoration found on ceramics of the Greek Classical period
(Bundrick 2008) and in the fresco scenes and stone vessel carvings of the Bronze Age
period on Crete (Trnka 2007) are examples of this category of evidence.

A second material type, which falls into this category of indirect sources, is the
loomweight. While loomweights cannot tell us specifically about the textiles, which they
were used to produce, they do, nonetheless, provide a number of clues as to the nature
of the weaving stage of the textile manufacturing process. This is particularly evident in
the case of both the prehistoric and historic periods on the island Crete (Burke 2010 pp.
50–1). For example, the weight of an individual loomweight can give us a strong
indication as to the strength of the textile fibres, on which it was employed.193 A general
rule holds that the heavier the loomweight, the heavier the associated fibre and by
inference the heavier and perhaps tougher the resultant textile (Burke 2007 p. 67). Using
this and other insights as a guide, we can begin to construct from loomweight
assemblages quite a nuanced picture of past textile production norms and activities. For
instance, were an archaeological context to contain a multitude of relatively light
loomweights, this would indicate that the location was once utilised for the production
of fine textiles perhaps made of silk or light wools. Conversely, the discovery of a
concentration of medium to heavy loomweights would suggest that much heavier
textiles, perhaps carpets, were being manufactured using fibres such as flax and/or
heavier wools.

192 Gordion, a substantial Phrygian settlement was home to a thriving textile industry during the
early part of the 1st Millennium BCE as evidenced by the archaeological recovery of large
installations dedicated to the manufacture of textiles (Burke 2010). The region is referenced by a
number of historical sources (Strabo Geog 1.3.21, Euseb. Chron 0.1.21.1) as being famous for its
wealth and this image was personified in the character of King Midas (Hdt. 1.14.35). It is likely
that the historical reports of Phrygia’s wealth were largely accurate and that it was derived from
the management of vast flocks of sheep and the manufacture and sale of textiles produced from
their wool.

193 On this basis, it follows that weight is really the only meaningful dimension when considering
loomweight material (Burke 2010 p. 117).
The loomweight form and their use within the production of textiles

Loomweights were used as an accessory in the manufacture of textiles. They were employed as a key component part of a device known as a warp-weighted loom. The earliest looms emerged in the Middle East region during the Neolithic and based on pictorial evidence we are able to deduce that these earliest incarnations were temporary in nature (Barber 1991 chap. 2). These ‘proto’ looms would have consisted of two beams of a rigid material, most likely wood. The weaver used these beams to stretch out a series of parallel threads spun from an organic material such as wool or flax. This is referred to as the warp. The weaver held these warp threads taut by pulling the two parallel beams apart using their feet and arms, as is depicted in the Han Dynasty bronze figurine shown in Figure 126. Using this method, textile widths were limited by the width of the weaver’s legs.

![Figure 126: 1st Millennium early Han Dynasty bronze figurine showing a weaver (Barber 1991 fig. 3.1)](image)

This temporary loom evolved into what is known as the horizontal ground loom, which employed the use of a more rigid rectilinear wooden frame and which was operated horizontally on a flat surface. We find this type of loom attested in Egyptian and Mesopotamian artwork at the very beginning of the Bronze Age.\(^{194}\) In the case of the horizontal ground loom, the warp threads are tied in a series of rows that run from one end of the loom to the other. The weft threads are interleaved with the warp threads and this interweaving process creates the textile.

\(^{194}\) A dish found at Badari in Egypt and a seal from Susa, both 4th Millennium, appear to depict the use of a ground-loom (Barber 1991 fig. 3.3, 3.4).
For these first two horizontal loom types, the loomweight is not required. It is only with the arrival of the warp-weighted loom, which first appeared during the late 5th or early 4th Millennium BCE in the Near East and somewhat later in the wider eastern Mediterranean region, that the loomweight is first attested.\(^{195}\) This new type of loom differs from the earlier horizontal variant in that it is orientated vertically. It is not exactly clear why this change came about but it is conjectured that it would have made the process of working the full surface of the textile far easier for the weaver (Barber 1991 pp. 81–2). In order to operate a warp-weighted loom, you must first lean it against a vertical support such as a wall. One end of each of the warp threads is then tied to the more elevated horizontal beam. The other end of each warp thread is tied to a loomweight and this serves to hold it vertically taut, as can be seen in the construction shown in Figure 127.\(^{196}\)

![Loom Reconstruction](image)

**Figure 127:** a loom reconstruction, Gordion Terrace Building (Burke 2007 fig. 10.6)

After the introduction of the warp-weighted loom, there is very little change of note to be found in loom design for many thousands of years and it is for this reason, that the loomweight is such a common category of find within the archaeological record at sites

---

\(^{195}\) It is difficult to accurately date the arrival of the warp-weighted loom due to the paucity of surviving evidence. Having said this, we can say that once loomweights begin to appear in the archaeological record, it is possible to speak with confidence of the use of the warp-weighted loom. The first attested loomweights appear in the Cretan archaeological record in the Late Neolithic (Stratum I) level at Knossos in the cuboid form (Burke 2010 p. 23).

\(^{196}\) Note that more than one warp thread can be attached to a single loomweight.
right across the southern European region from the Late Neolithic on through to the Bronze Age, Iron Age, Classical, Hellenistic and Roman periods.\textsuperscript{197}

Given its function, the only real requirement of the loomweight is to exert a certain weight onto one or more warp threads. This limited design specification resulted in there being only a small range of forms, which the loomweight could take. Clay is an ideal source material with which to make loomweights as its material characteristics give the maker a high level of control over the artefact’s shape and weight, which is crucial. A loomweight also needs to be given one or more holes with which to attach the warp threads. Besides these two fundamental requirements, however, there is very little else that a loomweight needs to do. Therefore, for the most part loomweight design over the millennia is relatively conservative and resistant to change. For example, on Crete we encounter predominantly discoid specimens during the Bronze Age and pyramidal during the Archaic and Classical periods, and within these periods there is very little need for and evidence of much variation in form.\textsuperscript{198}

It is this relative conservatism of form, when compared against other material categories (most obviously, pottery), which makes loomweights as a group much less useful when employed by the archaeologist as a dating tool. For instance, while a pottery specialist can with a relatively high degree of certainty assign a decade and location to a pottery sherd, an ancient textile specialist would be forced to talk in much broader temporal (200-300 years at best) and geospatial (regions rather than sites) terms when confronted with an analogous loomweight artefact.

\textsuperscript{197} Note that there is a question as to whether ceramic loomweights were used during the later Roman and Byzantine period on Crete and in Greece more generally (Davidson 1952 p. 147). It might have been the case, for example, that the use of ceramic loomweights was abandoned in favour of bags of sand or small stones from this point on.

\textsuperscript{198} Cretan and Greek loomweights do occasionally include decorative flourishes and some are even marked by incisions, which are likely to have served some administrative function (for example, at Palaikastro (Burke 2010 p. 43)) but these instances are quite rare and none are attested at Priniatikos Pyrgos.
Methodology

*Recording the Priniatikos Pyrgos loomweights*

Almost 200 individual loomweights or loomweight fragments were recovered during excavation at Priniatikos Pyrgos between 2005 and 2010. Besides being interesting as a material type for the reasons already outlined, the Priniatikos Pyrgos loomweights are also significant to this research project because they are the only category of find, whose study was conducted entirely by the author. Each step of the process, from empirical recording to digitisation and finally to publication to the Archaeological Semantic Web, was entirely under my own control. Therefore, many of the challenges (inconsistent use of vocabulary and recording methodologies) described in Chapter 5 encountered in the context of other material types (contexts, ceramic finds, et cetera) were avoided in the case of the Priniatikos Pyrgos loomweights.

For the most part, the initial recording of the loomweights followed a largely standard procedure observed by the majority of modern archaeological post-excavation projects (Wardle et al. 2001 sec. 6.7). The first step saw each loomweight being weighed to a precision of 1 gr using an electronic scale. An estimate was then made as to the percentage survival of the artefact.\(^{199}\) Each loomweight was then classified into one of eleven types, which are listed in Table 27.\(^{200}\)

<table>
<thead>
<tr>
<th>Type</th>
<th>Date range</th>
<th>Artefact count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greek pyramidal</td>
<td>480-323 BCE</td>
<td>85</td>
</tr>
<tr>
<td>Hellenistic pyramidal big</td>
<td>323-146 BCE</td>
<td>1</td>
</tr>
<tr>
<td>Greek pyramidal wedge</td>
<td>480-323 BCE</td>
<td>4</td>
</tr>
<tr>
<td>Greek pyramidal broad base</td>
<td>480-323 BCE</td>
<td>6</td>
</tr>
</tbody>
</table>

\(^{199}\) While arriving at an estimate of this type is undoubtedly subjective, it was, nonetheless, necessary to include this step as about half of the entire Priniatikos Pyrgos assemblage was found to be in a partial state of survival. Also, given that a loomweight’s primary meaningful dimension is weight, a partial weight recording without a statement as to how this equates to the whole, would be of little subsequent analytical value.

\(^{200}\) These type names included here have been defined by the author. No canonical vocabulary exists for loomweights, as is the case with Roman coins (cf. *Roman Imperial Coins* volumes) or for certain categories of ceramic data (Yon 1976, 1982), and to a large extent loomweight scholars use their own typologies that are based almost exclusively on form. While a number of these labels might correspond to those used by the classification systems of other archaeological projects, this typology should only be used to identify material from the Priniatikos Pyrgos assemblage.

Of which there were 175 loomweights found in total.
Table 27: Priniatikos Pyrgos loomweight type groupings and their artefact counts

<table>
<thead>
<tr>
<th>Type</th>
<th>Date Range</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minoan spherical</td>
<td>1750-1405 BCE</td>
<td>3</td>
</tr>
<tr>
<td>Roman conical</td>
<td>146 BCE - 330 CE</td>
<td>4</td>
</tr>
<tr>
<td>Minoan discoid</td>
<td>2650-1100 BCE</td>
<td>24</td>
</tr>
<tr>
<td>Minoan discoid flat-top</td>
<td>2650-1100 BCE</td>
<td>21</td>
</tr>
<tr>
<td>Hellenistic discoid</td>
<td>323-146 BCE</td>
<td>14</td>
</tr>
<tr>
<td>Discoid pendant</td>
<td>unknown</td>
<td>2</td>
</tr>
<tr>
<td>Hellenistic biconical</td>
<td>323-146 BCE</td>
<td>10</td>
</tr>
<tr>
<td>Unknown</td>
<td>unknown</td>
<td>1</td>
</tr>
</tbody>
</table>

By analysing comparanda from the sites of Knossos (Alberti 2008; Popham 1984), Palaikastro, (Burke 2010 chap. 2), Athens (Davidson et al. 1943), Corinth (Davidson 1952) and Azoria (Stefanakis et al. 2007 pp. 2003–2004) and taking into account the dates given for the ceramic material found in association with these artefacts, a date range was assigned to each loomweight type. These date ranges, as well as the artefact counts for each grouping, are also included in Table 27 and in the pie chart shown in Figure 128.

![Figure 128: a distribution of loomweight types found at Priniatikos Pyrgos](image)

Each artefact was then photographed a number of times to publication standards. Each photographic set includes a photo emphasising the subject’s diagnostic aspects with and
without accompanying scale and identification label. A composite image, which includes a sample of each of the loomweight types found at the site, is reproduced in Figure 129.

![Figure 129: the range of loomweight types identified at Priniatikos Pyrgos](image)

Lastly, in order to create a virtual study assemblage for the material, a sample number of loomweights was recorded photogrammetrically as a textured 3D object using the iOS app, Autodesk 123D Catch (Autodesk, Inc. 2015). A screenshot of a Roman biconical example is shown in Figure 130.

![Figure 130: screenshot of a Priniatikos Pyrgos biconical loomweight taken in Autodesk 123D Catch](image)
Following the practice established for all non-ceramic artefacts marked out for cataloguing at Priniatikos Pyrgos, the loomweights were entered into the catalogued objects table in the FileMaker database. Each of these entries came to be represented as an RDF resource of type la_pp_ont:LA_E6_CataloguedObject on linkedarc.net. As such, each set of RDF triples relating to a particular loomweight artefact included the same basic information (name, description, set of empirical measurements, associations with the contexts or loci in which it was found, et cetera), as was the case for all catalogued objects. Again in keeping with the practice used for other catalogued items, each loomweight’s type was also included in this RDF representation and a set of vocabulary entries was created to realise the loomweight typology described already. These values were linked to each loomweight subject using the ecrm:P2_has_type predicate, in keeping with standard CIDOC CRM practice. Lastly each loomweight’s estimated creation date was set by linking a Priniatikos Pyrgos RDF period URI or range of period URIs to the loomweight’s subject via the ecrm:P108i_was_produced_by predicate. By way of example, the RDF representation for the loomweight ‘05-001’ is reproduced in graph form in Figure 131.

Figure 131: an RDF representing the loomweight 05-1035
Querying the loomweight data using SPARQL

The objective of the PPTextiles project was to create a map-based view onto the loomweight assemblage at Priniatikos Pyrgos. The hope was that this perspective would show up changing patterns of the use of space at the site through time. In order to access this perspective, it was first necessary to find the geospatial location of each loomweight, its loomweight type and the period assigned to it. As has been argued in the first case study in this chapter, SPARQL is by far the most flexible way of accessing this type of specific information from an Archaeological Semantic Web resource. For example, the SPARQL query shown in Table 28 asks the linkedarc.net server for the subject URIs, the names and the geolocations of all the discoid loomweight artefacts found at the site.

<table>
<thead>
<tr>
<th>Table 28: finding Priniatikos Pyrgos's discoid loomweights using SPARQL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>12</td>
</tr>
<tr>
<td>13</td>
</tr>
<tr>
<td>14</td>
</tr>
<tr>
<td>15</td>
</tr>
<tr>
<td>16</td>
</tr>
</tbody>
</table>

In turn, the data might also be interrogated from a chronological point of view. The SPARQL query shown in Table 29 asks the server for all of the loomweight artefacts, which were created and used during the Classical period.

<table>
<thead>
<tr>
<th>Table 29: finding Priniatikos Pyrgos's Classical loomweights using SPARQL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
</tbody>
</table>
The results of this type of query could be presented in a number of different ways. For instance, Gephi could be used to create a graph visualisation of the data following the lead taken by the Priniatikos Pyrgos Trench 2 stratigraphy case study. The ‘Yifan Hu Proportional’ layout could be employed to highlight the periods, which were associated with the most loomweight finds, for instance. Or alternatively, we could use Gephi’s GeoLayout plugin (Bastian 2013) to present a spatial distribution analysis of the loomweight assemblage.

Both of these options would provide their own particular insights into the cloud of Priniatikos Pyrgos loomweight data. For this project, however, I chose to write a small and specialised web app, which would send tailored queries to the linkedarc.net SPARQL endpoint as and when the user requests an update of the visualisation. The reason that I chose to use this particular workflow was to highlight just how easy it is for a third party application such as the PPTextiles web app to access content that is made available as RDF.\(^{202}\)

**The PPTextiles web app**

The PPTextiles web app can be accessed using a web browser at http://pptextiles.linkedarc.net. While it exists on a subdomain of linkedarc.net, its architecture is entirely separate to the linkedarc.net server and web app. Like the linkedarc.net web app, it consumes the services of the linkedarc.net server by sending queries to the server’s SPARQL endpoint.

---

\(^{202}\) To date, this model, which sees a web app consume the services of an Archaeological Semantic Web resource other than itself, has been seen very infrequently in the cultural heritage app space. There are, however, some exceptions to this rule. For instance, the Seneschal project have made available a number of so-called ‘widgets’, which allow data input interface designers to include access to the Seneschal RDF vocabulary resources in real time (Binding 2015). WordPress users can also access Pleiades resources in a similar way using the ‘Ancient World Linked Data’ plugin (Jones 2012). Latterly, projects such as Europeana have been investigating whether or not the Web API is a service that humanities researchers are really interested in exploiting (Edmond & Garnett 2015).
The web app’s functionality is relatively simple. The user is first presented with a plan of the Priniatikos Pyrgos site (Figure 132). This plan includes contour lines indicating the site’s topography, trench demarcations and the major architectural elements contained within each trench. The user is able to manipulate the map view by zooming in and out and panning about as required. The plan layer itself is georectified onto an underlying world map. The web app’s mapping function is delivered using the Leaflet JavaScript library (Derrough 2013) and the map tiles for the site plan were constructed using the MapTiler application (Klokan Technologies GmbH 2015).

---

203 This plan image is derived from a version generated using the Quantum GIS application by the project’s GIS specialist, William Megarry.

204 The code for the PPTextiles project is available as Open-source Software at https://bitbucket.org/flynam/linkedarc-pptextiles.
By default the map is marked with the locations of all of the loomweight artefacts (Figure 133). These positions are displayed as a heatmap, with areas of higher density being rendered as a red colour. As the density falls, the colour changes from red to yellow to green to blue until the heatmap layer disappears in areas where there is a total absence of loomweights. At this stage, it is difficult to see a clear pattern for the loomweight finds, which is unsurprising given that we are looking at the spatial spread of the loomweights across all periods. We will need to delve more deeply into the data and to consider it from a typological or chronological point of view, if we are to begin to discern its meaningful patterns.

Figure 134: filtering the loomweight distribution by type

In the top-right-hand corner of the map view, three controls allow the user to filter the data, which is retrieved from the server. The ‘Loomweight types’ dropdown lists all of the loomweight types (Figure 134). One or more of these types can be selected and by clicking the refresh button to the left of this dropdown a fresh call will be sent to the SPARQL server with this information included in the query. When the server responds, it displays an information box summarising the results (Figure 135). The number of artefacts found to match the query’s criteria is noted (in the case of the discoid
loomweights, the figure is 18) alongside the SPARQL query, which was sent to the linkedarc.net server.

Figure 136: the distribution of discoid loomweight finds at Priniatikos Pyrgos

Figure 136 shows a distribution map for all the discoid loomweights found at the site. This again does not present any entirely definitive spatial patterns, as Trenches 2, 3, 4 and Area G are all shown to contain loomweights of this type. It is also possible to query the loomweight assemblage on the basis of the dates of the artefacts. Let us consider whether adopting this approach yields any more telling results. Selecting to view only those loomweights dated to the Bronze Age produces the distribution map shown in Figure 137 with 30 specimens in total attested. While the majority of the trenches contain some Bronze Age loomweights, there certainly appears to be a concentration of finds in Trench 3 and Areas H and G on the western edge of the site. This general pattern tallies with the discovery of a kiln structure in the extreme northwestern corner of Area G (Hayden 2014 p. 16). Could this part of the site have served a predominantly industrial function manifested in the form of ceramic firing activity as well as textile production? The short answer is, possibly but more information is needed to confidently make such a statement. The observation of this Bronze Age loomweight concentration serves as a starting point with which to begin a process of investigation. The next step required would be to interrogate the contexts in which these loomweights were found.
When you zoom closer into the plan, the heat map view is replaced with a series of blue dots, which indicate a find spot for either a single loomweight or a group of loomweight found together (Figure 138). Clicking on a find spot presents a popup display. This contains a list of all of the loomweight artefacts found in that spot. These identifiers are also hyperlinks, which when clicked will bring you to the resource’s representation on
linkedarc.net. And from there the artefact itself and its context can be studied in greater detail.

Figure 139: the distribution of the Archaic and Classical loomweights found at Priniatikos Pyrgos

Moving further on in time, a slightly different picture emerges with the distribution of the 47 mainly pyramidal loomweights of the Archaic and Classical periods (Figure 139). These artefacts are relatively less common in the Area H/Trench 3 zone when compared to the Bronze Age data and there appears to be a slight increase in their numbers in Area G and in Trenches 2 and 4. Also, while small in quantity, there is now a loomweight presence in the Area A zone, which was entirely absent in the Bronze Age period.

Figure 140: the distribution of the Hellenistic and Roman loomweights found at Priniatikos Pyrgos
Lastly, the Hellenistic and Roman periods see a drop in overall figures, with only 14 loomweights in total being attested (Figure 140). There is also a general movement away from the contexts of the western edge of the site in favour of those of the central zone occupied by Trench 2, with the exception of a very localised deposit of two loomweights, 05-1093 (pyramidal-big) and 05-1111 (conical), which were found in Trench G2000. The overall amounts are low, however, and any interpretations of trends would be statistically unsafe due to this low sample.

**Research findings**

PPTextiles is an example of a lightweight web app, which consumes the services of the linkedarc.net server. In contrast to the linkedarc.net web app, it targets a particular subset of the data contained within the Priniatikos Pyrgos dataset. Its objective is targeted: to display distribution plans of the loomweight finds at the site and by doing so to present a narrative of the changing patterns of textile production through time.

The views presented by the PPTextiles web app tell us that the density of loomweight finds changes across time at Priniatikos Pyrgos. In the Bronze Age, most loomweights were deposited on the NW slope of the site in the regions of Trenches 3 and Area H. For the Archaic and Classical periods, this deposition pattern moves inland with an increase in loomweights being observed in Area G and Trenches 2 and 4. This is valuable information that can be used to build a series of hypotheses about the changing nature of the site's form and function across time.

Methodologically speaking, PPTextiles is an important example of how data re-use is encouraged and supported by Linked Open Data research environments and, more specifically, by the making available of these resources using a SPARQL interface. The PPTextiles web app development took about a week for one person to complete, which by most measures is a remarkably short period of time for the completion of a project such as this. Because the Priniatikos Pyrgos data has been structured using the documented CIDOC CRM and CRM-EH ontologies, the web app knows how to construct the SPARQL queries, which will be used to access the loomweight material. The queries do not just target the loomweight entities in themselves – PPTextiles also requires information about their geolocations. Therefore, information about the contexts or loci in which the loomweights are found is also requested. This multi-targeted
approach, while somewhat more convoluted, shows the power of SPARQL as it allows for the querying of multiple data types from within the one query.

The PPTextiles web app could be expanded upon to provide additional functionality by mining data from other Archaeological Semantic Web resources. It might for instance, draw comparisons between the loomweight artefacts found at Priniatikos Pyrgos with those stored in the British Museum. It could do this by exploiting the fact that the <http://linkedarc.net/vocabs/object-loomweight> resource, which is the type of all of the loomweight entities in the Priniatikos Pyrgos dataset, is a SKOS:Concept entity. Querying the <http://linkedarc.net/vocabs/object-loomweight> resource will show that it is considered to be the same basic concept as the British Museum’s vocabulary concept, <http://collection.britishmuseum.org/id/thesauri/x7732>. Knowing this, it is possible to construct a SPARQL query, which asks the British Museum’s SPARQL endpoint to return information about its loomweight collection. The query shown in Table 30 requests the find spots of these resources. All non-Greek artefacts could then be filtered out of the results, thereby making them more applicable to the Priniatikos Pyrgos assemblage.

```
SELECT ?s ?findPlaceLabel WHERE {
  ?s ecrm:P12i_was_present_at ?findDiscoveryEvent .
  ?findDiscoveryEvent ecrm:P7_took_place_at ?findDiscoveryEventPlace .
  ?findDiscoveryEventPlace skos:prefLabel ?findPlaceLabel .
  ?findDiscoveryEventPlace skos:broader ?placeBroader1 .
  ?placeBroader1 skos:prefLabel ?findBroader1Label .
  FILTER(REGEX(?findBroader1Label, '(crete)|(greece)', 'i'))
}
```

Table 30: locating the Greek loomweights in the British Museum using SPARQL

Similarly, the Archaeological Data Service’s SPARQL endpoint could be queried for resources that are of type <http://purl.org/heritagedata/schemes/mda_obj/concepts/97042>, which is also linked to the <http://linkedarc.net/vocabs/object-loomweight> linkedarc.net resource. The type of aggregating functionality described here, where one algorithm queries a first SPARQL endpoint and then another, is a perfect illustration of the

---

205 The <http://linkedarc.net/vocabs/object-loomweight> resource is linked to the <http://collection.britishmuseum.org/resource/thes/x7732> resource via the skos:exactMatch predicate.
Archaeological Semantic Web in action using SKOS as the semantic glue with which to bring all of these disparate resources together.

**Interpreting the British Museum cuneiform collection**

<linkedarc.net resource>Interactive outputs of the British Museum cuneiform case study</linkedarc.net resource> http://linkedarc.net/datamining/bmcuneiform

**Context**

The cuneiform system of writing has held a curious mystique for ancient world scholars ever since it first entered the Western consciousness in the accounts of early Near Eastern explorers such as Carsten Niebuhr, who during the late 18th century visited the site of Persepolis, which was then a part of Persia (Pollock & Bernbeck 2004 p. 7). Alongside the pyramids in Egypt and the megalithic complexes of northwestern Europe, the alien angularity of the cuneiform system has come to epitomise the great achievements of some of the earliest human endeavours. While once cuneiform and these other artefacts highlighted how little we knew of the antique world, they now signify just how much progress has been made over the last two centuries in our understanding of these periods.²⁰⁶

The earliest cuneiform documents evolved out of simpler pictogram-based systems and these for the most part were written onto soft clay canvases using a reed stylus to create the impressions in the clay. Clay was an enormously abundant and, therefore, cheap commodity for the people of the first Near Eastern settlements (Pollock 1999 p. 79).²⁰⁷ They used it build their houses, their temples and as an integral part of their bureaucratic system in the form of seals first and then counter artefacts (Pollock 1999 pp. 110–111).

²⁰⁶ Cuneiform, more than many other forms of material culture, also symbolizes the role that colonialism played in the development of early archaeology in the Near Eastern region. For more on this topic, consult Gosden (2004) and Trigger (1984) for a critical overview of the subject in general and Bahrani (1998) for a more specific look at the role of colonialism in Near Eastern archaeology.

²⁰⁷ Clay is a component ingredient of soil alongside water, organic material, silt and sand. It is estimated that soil makes up about ¾ of the earth’s crust (Facey 1997 p. 10). Archaeologists working at Neolithic and Bronze Age tell sites in the Middle East often observe large depressions on the outer edges of these sites. It is speculated that these depressions would have served as sources of clay for the inhabitants of the settlement (Wilkinson 2003 pp. 125–126).
As such, it was a natural progression for these peoples to use the same material as a medium on which to write and to create documents. It was the physical characteristics of the particular type of reed, which was also found in plentiful supply growing along the banks of the great Euphrates and Tigris Rivers, that when pressed into the water-softened surface of a small pillow-shaped clay tablet produced the distinctive wedge-shaped incisions that we know today as cuneiform.

As with most early forms of writing, these first cuneiform documents were used exclusively for administrative purposes, as records of exchange activity (Mieroop 2005 pp. 13–17; Nissen 1986). The consensus academic view is that these proto-civilisations emerged first as temple economies. This model saw an elite class administer the social and economic lives of a dependant population through the material and metaphysical presence of the temple (Pollock 1999 p. 93). In times of plenty the state acted as a repository for the surplus products of what was almost exclusively a subsistence agrarian economy, while at the same time taking a proportion of the output as tribute and redistributing the remainder as rations to the people. Then, whenever fortunes took a turn for the worse, as they inevitably did (Rosen 2007), the temple stores acted as a buffer against the social hardships that would otherwise come about as a result. As a by-product of this system, society gradually came to be stratified as certain individuals and groups came to accumulate and to pass on wealth intergenerationally (Westbrook 2003 sec. 6.3). As advances in technology and social cohesiveness came about, temple communities evolved into city-states and by this point the earlier management systems, which were employed by the proto-economies, exceeded the capabilities of human cognition alone. A consequence of this was that the city-state administrators needed to

---

209 The same is true in the case of the Mycelean script Linear B and from what we can tell of the Linear A script, it was also true for the Minoan script.
210 There has been much debate (Finley 1973; Scheidel & Reden 2002) as to whether it is anachronous to refer to ancient systems of production and consumption as economies given all of the modern connotations that comes with the use of this word. Scholars even argue about whether the term is acceptable in the context of the Roman Empire (Bingen 2007; Greene 1986; Temin 2006) and it is clear that the scales of the earliest Mesopotamian civilizations are entirely more diminutive than this.
211 Another artefact type associated with this redistributive system in the Mesopotamian states is the Bevel-Rimmed Bowl, which were produced in their millions as disposable vessels for the dishing out of rations (Johnson 1973).
212 For example, the self-sowing plough brought about enormous gains in the production efficiency of land (Oates 1986 pp. 193–194). Advances in irrigation system technology had similar positive impacts (Wilkinson 2003 pp. 71–99).
externalise their accounting systems and the cuneiform tablet became the ultimate manifestation of this drive towards material recording practice.\textsuperscript{213}

Once the Mesopotamian scribes got to grips with the new system, it exploded in popularity across the region with almost blanket adoption occurring possibly by the end of the 4\textsuperscript{th} Millennium and certainly by the beginning of the 3\textsuperscript{rd} Millennium (Foster 2007 p. 261). City-states were producing literally thousands of these documents and storing them in restricted archive installations. With time the subject matter of these records expanded, as well. What we would now term literary texts started to emerge with certainty by the mid-3\textsuperscript{rd} Millennium at Fara and Abu Salabikh. These were written in the Sumerian language (Holm 2007 p. 269). These first non-administrative texts dealt with a large number of subjects\textsuperscript{214} but prominence was always reserved for themes dealing with matters of divinity, of which there were many (Foster 2007 p. 262). One of the great advantages of the cuneiform system was that, like the Latin alphabet, it could be used to render more than one language. This proved to be an absolute necessity in a region, which hosted such a huge range of peoples and tongues, some of which we know about today (for example, the aforementioned Sumerian, which is the earliest language to be committed to text, Akkadian, Egyptian, Hittite, Elamite, Hurrian, Urartian) and inevitably many others that are beyond our current awareness or comprehension (Rubio 2007 p. 79).

This decision by the populations of Mesopotamian city-states in the beginning of the Bronze Age period to use clay tablets as a means of externalising and storing vast amounts of information relating to a whole range of activities and belief systems has had a profound effect on the ways that modern scholars are able to study the lives of these groups of people from a remove of several thousand years. As with many assemblages that come to be excavated as part of an archaeological investigation, the clay tablets that survived this vast expanse of time sealed within the archaeological record have managed to do so as a result of a disaster having befallen them. Fire serves to vitrify the otherwise porous and fragile surface of the clay and this essentially freezes the imprint of the scribe’s pen upon the tablet. Luckily, these incidents, whether intentional or not, were

\textsuperscript{213} This concept of external memory is first proposed by Donald (1991) as one part (the other being theory creation) of the third stage in his development of the modern human mind.
\textsuperscript{214} Holm (2007) outlines a categorisation of Near Eastern genres as follows: mythological narratives, epics, laments and prayers, didactic literature, autobiographies and love poetry.
quite common relative to the vast time frames involved and teams of archaeologists have managed to unearth and preserve by conservation many hundreds of thousands of these artefacts originating from a whole range of sites across the region (Earl et al. 2011 p. 152).

**Research question**

While the majority of these cuneiform artefacts are now found in the museum collections of Iran, Iraq, Syria and the other states of the region, many more reside in public and private collections across the globe. The British Museum has gathered together one of the world’s largest cuneiform collections, claiming to have in the order of 130,000 texts or text fragments, ranging in date from the Early Dynastic through to the Neo-Babylonian periods (Trustees of the British Museum 2015). This case study profiles the British Museum’s cuneiform collection using a workflow that is built primarily upon Archaeological Semantic Web practices. It looks to contextualise the biographies of the cuneiform artefacts contained within the museum’s collection and to a lesser extent it explores their inscriptions as textual artefacts.

**Methodology**

**Getting a sense of the British Museum’s cuneiform Linked Open Data**

As was explained in Chapter 4, the British Museum has played a leading role in the development and adoption of the CIDOC CRM model. As a result, while it includes a number of proprietary extensions in its implementation of the CRM’s classes, for the most part the museum’s data is structured as is prescribed by the CRM guidelines (Le Boeuf et al. 2015; Oldman & Rahtz 2014). As such, it represents an ideal opportunity on which to consider the process of querying an initially unknown RDF triplestore, which is structured using the known CRM. These queries will be run through the British Museum’s SPARQL endpoint.

We need to begin by finding the object types that represent the museum items, which include cuneiform inscriptions. It would be sensible, therefore, to start by looking for the RDF entities, which represent the museum’s artefact assemblage. Following the CRM’s guidelines, we need to ask the British Museum’s SPARQL endpoint about its

---

215 These extensions have a base address of <http://collection.britishmuseum.org/id/ontology/>.
ecrm:E22_Man-Made_Object instances and we can achieve this using the SPARQL query shown in Table 31.216

```
1 SELECT (COUNT(DISTINCT ?artefact) AS ?artefactCount) WHERE {
2     ?artefact a ecrm:E22_Man-Made_Object
3 }
```

Table 31: counting up all of the artefacts in the British Museum using SPARQL

Running this query returns a value of 2,062,303 ecrm:E22_Man-Made_Object resources. We now need to narrow this list down to the artefacts, which also have inscriptions. The CRM has a predicate ecrm:P65_shows_visual_item, which associates a ecrm:E22_Man-Made_Object instance with a visual representation modelled as a ecrm:E36_Visual_Item instance. An inscription, being a visual representation, would appear to fit this description. The SPARQL query shown in Table 32 will ask for all ecrm:E22_Man-Made_Object instances, which link to a ecrm:E36_Visual_Item instance via the ecrm:P65_shows_visual_item predicate.217

```
1 SELECT ?artefact ?visualRep WHERE {
2     ?artefact a ecrm:E22_Man-Made_Object .
4 } LIMIT 100
```

Table 32: retrieving the first 100 British Museum artefacts with visual representations using SPARQL

As you can see, none of the URI prefixes have been included in the header of this SPARQL query. The British Museum’s SPARQL query interpreter assumes when a prefix is not included that it points to a standard namespace. For example, ecrm points to the most commonly used CIDOC CRM namespace, <http://erlangen-crm.org/current/>.

Note that at this point, as we are in the process of learning about the museum’s data structure, we can limit this search to the first 100 instances that match our criteria. Later it will become necessary to ask for a total view of all matching entities.
When we run this query, we get a list of visual representations but not all are inscriptions. While we were expecting instances of type ecrm:E36_Visual_Item, we actually receive a list that includes ecrm:E38_Image, ecrm:E72_Legal_Object and a number of other types as well. What is going on here? Is this perhaps ontological sub-classing at work? To find out more, let us consult the CRM’s guidelines for the use of the ecrm:P65_shows_visual_item predicate (Figure 141).

These notes tell us that the ecrm:P65_shows_visual_item predicate has a domain (i.e. the permissible subject type) of ecrm:E24_Physical_Man-Made_Thing and a range (i.e. the permissible object type) of ecrm:E36_Visual_Item. Perhaps, the ecrm:E38_Image and ecrm:E72_Legal_Object classes are sub-classes of ecrm:E36_Visual_Item and this would render their use as an object in a triple including the predicate ecrm:P65_shows_visual_item legal. The CRM class hierarchy (Figure 142) clearly shows, however, that while ecrm:E38_Image is a sub-class of ecrm:E36_Visual_Item, ecrm:E72_Legal_Object is not. In fact, ecrm:E72_Legal_Object is a super-class of ecrm:E36_Visual_Item!
The only remaining explanation for this apparent anomaly is that the CRM has been incorrectly implemented by the British Museum in this case. While an annoyance, this type of situation can be encountered quite frequently when querying resources on the Archaeological Semantic Web. Technically, it is allowed for by the ‘loose typing’ of RDF. This means that while a structuring ontology (in this case, the CRM) prescribes a set of rules that should be followed, there is no technical imperative on the RDF creator to do so. And when mapping a complex dataset such as the British Museum’s cuneiform collection onto an equally complex ontology such as the CRM, it is inevitable that the occasional ontological incompatibility is introduced in error.

Let us return now to our research question. ecrm:E34_Inscription instances are allowed to be linked to a ecrm:E22_Man-Made_Object instance via ecrm:P65_shows_visual_item as we have explained above. Let us narrow the query down again so that we only receive a list of the artefacts, which have a link to an inscription instance (Table 33). This will exclude all of the visual representations linked to the resource that we are not interested in.

```
1 SELECT DISTINCT ?artefact WHERE {
2   ?artefact a ecrm:E22_Man-Made_Object .
4   ?visualRep a ecrm:E34_Inscription
5 } LIMIT 100
```

Table 33: retrieving the first 100 British Museum artefacts with inscriptions using SPARQL

This returns a list of artefacts, which have inscriptions associated with them. If we look at the human-readable representation of one of these inscription resources (for example, the ‘cartouche of Malonaqan’ inscription shown in Figure 143), we can see that it lists a number of CRM predicates. This is to be expected. However, it also includes quite a number of the British Museum’s custom ontology predicates, which on the basis of their labels, would appear to provide valuable information sources about the museum’s cuneiform’s artefacts.

---

The hyperlinks for these predicates can be followed to get access to their domains, ranges and descriptions.
The bmo:PX_inscription_script predicate provides a link to a skos:Concept instance, which represents a vocabulary entry for one of the many script types found in the museum’s collection. One of these is the <http://collection.britishmuseum.org/id/thesauri/script/cuneiform> resource, which represents the thesaurus entry for the cuneiform script and we can inject this additional criterion into our SPARQL query as is shown in Table 34.

```
1 | SELECT DISTINCT ?artefact WHERE {
2 | ?artefact a ecrm:E22_Man-Made_Object .
4 | ?visualRep a ecrm:E34_Inscription .
6 | ) LIMIT 100
7 | }
```

Table 34: retrieving the first 100 British Museum artefacts with cuneiform inscriptions using SPARQL

Following a similar workflow, we can augment the search with criteria, which ask for the artefact’s find place, the person associated with the find, the discoverer’s nationality and gender as is shown in Table 35.

```
1 | SELECT DISTINCT ?artefact ?placeName ?discovererName ?discovererNationalityName ?discovererGender WHERE {
2 | ?artefact a ecrm:E22_Man-Made_Object .
4 | ?visualRep a ecrm:E34_Inscription .
6 | ?visualRep ecrm:P3_has_note ?inscriptionText .
7 | ?artefact ecrm:P12i_was_present_at ?findDiscovery .
8 | ?findDiscovery ecrm:P7_took_place_at ?discoveryPlace .
9 | ?discoveryPlace skos:prefLabel ?placeName .
10 |)
11 |}
```

Table 35: retrieving the first 100 British Museum artefacts with cuneiform inscriptions and additional criteria using SPARQL
Cleaning the data

Running this last query returns a list of 16,176 entries and a sample of this data as seen in Excel is reproduced in Figure 144. For the most part, this is good quality data that can be used to produce interesting interpretations. Certain facets of it, however, require some cleaning, most notably in the placeName values. These values identify the places associated with the discovery of each artefact. The majority of these data will have started out life as an entry in one of the museum’s numerous artefact catalogues, which date back to when the museum first opened its doors to the public in 1753 (Moser 2006; Wilson 2002). Artefacts bearing cuneiform inscriptions would have first arrived into its stores not long after this when by the late 18th century the museum was sponsoring large-scale excavation projects in the Near East as part of its efforts to position itself as one of the great collections in the world (Eisler 1996; Gibson 2008). As a result, many of the place names that appear in the results refer either to names used in the 18th or 19th centuries or to ancient place names.

Normally, place name data such as this can be enriched by requesting its longitude and latitude coordinates from a geocoding service, such as Google Maps Geocoding API (Ge 2005). Google Maps is extremely proficient at retrieving this type of information for...
modern place names but ancient place names are another matter entirely. 58% of these place names were not recognised by Google’s geocoding service. Therefore, the only remaining alternative was to find and input this information manually. While the list of 16,176 entries referenced in total just 62 unique place names, which equated to 48 unique locations, it still took a number of hours to find and enter these geo-coordinates by hand. This should serve as a reminder of the efficiencies delivered when digital automation is applicable and available.

Mapping the collection’s cuneiform

Using the geo-coordinate information derived during the data cleaning stage, it was possible to construct the distribution map of the 48 locations shown in Figure 145. In a single image, this map summarises over 200 years of cuneiform research at the museum. The sites are marked by colour-coded and differentially sized pins, which indicate the relative importance of each of the sites in terms of their cuneiform assemblages. 38 of the 48 sites yielded less than 100 individual cuneiform inscriptions for the museum and the majority of the entire collection is made up of inscriptions originating from the sites of Nineveh, of which there are 30,986 inscriptions, and Sippar, which is associated with 29,426 items.

Figure 146: a popup window providing more information about the site of Nineveh on the cuneiform distribution map

---

219 As of 18 June 2015. This test was facilitated using the Open Refine application.
220 The equivalent operation performed using a Google Reverse Geocoding setup, would be completed in a matter of seconds.
For the case study, it was decided to store the distribution map data in a Google Fusion table (Gonzalez et al. 2010) and Google provides a JavaScript API that can be used to access this content. The web app, which displays the distribution map, has also been engineered to provide a small degree of interactive functionality. Clicking on any of the site markers presents a popup window displaying the site’s name and its artefact count (Figure 146). It would be relatively straightforward to expand this functionality in a number of different ways, for instance by providing links back to the source data on the British Museum’s server.

**Focussing in on the contributors**

![Figure 147: the British Museum’s cuneiform inscription collection by discovering nation](http://linkedarc.net/datamining/bmcuneiform)

While the distribution map shown in Figure 145 tells the reader about the geographical origins of these artefacts, it does not tell them anything about the people involved in their discovery. We already have the nationalities for each of these individuals and the chart shown in Figure 147 summarises this information. Unsurprisingly, we can see that British nationals were chiefly responsible for the running of the large-scale projects, which generated the lion’s share of the museum’s cuneiform collection. Iraqi nationals are next in the list but at 1,104 inscriptions, they are markedly less than the 30,203 inscriptions for the British. German (156 inscriptions) and Syrian (72 individuals)

---

221 There is no reason why this data could not be stored in a number of other datastore types, be they Linked Open Data compliant or not. Google Fusion was chosen in this case because of its close integration with the Google Maps API.
222 This web app is available at [http://linkedarc.net/datamining/bmcuneiform](http://linkedarc.net/datamining/bmcuneiform).
223 This and the following series of charts were created using the online tool, Plotly (n.d.).
individuals come next with only a handful of inscriptions been submitted by French and Americans.

Let us look again at the data and consider the contributors themselves. The dataset that we mined from the British Museum’s SPARQL endpoint contains a contributor’s name and a breakdown of the top ten contributors by inscription count is shown in Figure 148. Included in this list are luminaries of Mesopotamian archaeology. Figures such as Layard (Fagan 2003 pp. 51–54; Holloway 2003; Larsen 2009), Woolley (Fagan 2003 pp. 117–120) and Smith (Hoberman 1983) shaped the core narratives that continue to structure the discipline to this very day. Hormuzd Rassam tops the group with a total of 37,699 or 53% of the total inscriptions. Unknown to many outside of Assyriology circles, Rassam played a leading but controversial role in the early development of the field (Crossen 2011 p. 462). He began his archaeological career working under Layard and when Layard moved on to pastures new he continued working alone excavating at sites such as Nimrud and Nineveh (Reade 1993). Were it not for the fact that Rassam became a citizen of the British Empire, the breakdown of contributing nations shown in Figure 147 would be very different and this shows an obvious problem with data that is designed to contain only a single value. Would it not be more meaningful to list two nationalities for Rassam? After all, he spent much of his life as an Iraqi citizen.

In total there are just 34 unique names listed as contributors for all of the artefacts in the collection. This small return is disappointing given the enormous effort that must have
gone into the excavation and conservation of these objects. Where are the names and biographies of all of these other individuals? Documentary evidence of these ‘subaltern’ archaeologists tend to be sidelined in the pursuit of the field’s grand narratives, with these being concerned almost exclusively on the roles of the principal investigators (Currie 1995; Guha 1982). While regrettable, for an investigation to be defined by the limitations of one’s source material is no new phenomenon of the digital age; archaeologists have always had to frame their interpretations based on the materials at their disposal.

While noting these shortcomings, we can, nonetheless, flesh out the names that we do have by sending a series of additional queries to the British Museum’s SPARQL endpoint. The query shown in Table 36 asks the British Museum for the birth and death dates of ‘Dr Harry Reginald Holland Hall’ and for any biographical notes that might also be in their triplestore.\(^\text{224}\)

\begin{verbatim}
SELECT DISTINCT ?discoverer ?dateBirth ?dateDeath ?note WHERE {
  ?discoverer a ecrm:E21_Person .
  ?discoverer skos:prefLabel 'Dr Harry Reginald Holland Hall' .
  ?discoverer ecrm:P98i_was_born ?birth .
  ?birthTimespan ecrm:P82a_begin_of_the_begin ?dateBirth .
  OPTIONAL { ?discoverer ecrm:P100_died_in ?death .
                ?deathTimespan ecrm:P82a_begin_of_the_begin ?dateDeath } .
  OPTIONAL { ?discoverer ecrm:P3_has_note ?note } }
\end{verbatim}

Table 36: getting the birth and death dates of Dr Hall from the British Museum using SPARQL

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure149.png}
\caption{a timeline of the British Museum’s cuneiform inscription contributors}
\end{figure}

\(\text{}^{224}\) In practice, this series of SPARQL queries (one for each of the 34 individuals returned by the previous query) can be constructed procedurally for each of the individuals’ names using an application such as Open Refine.
We can now use this new information to construct a timeline (Figure 149), showing the personalities involved in the creation of the cuneiform collection at the museum. This interactive timeline was created in a matter of minutes using Timeline JS (Northwestern University 2013), an excellent online service created by Northwestern University’s innovative knight lab project (Northwestern University 2015). To create a Timeline JS visualisation, you must first populate a Google Docs spreadsheet template with rows of time-based information such as that which is contained in our dataset about the British Museum’s cuneiform contributors. The spreadsheet must then be published to the web as an Open Access document. The Timeline JS library then accesses the document using its unique identifier and pulls out information such as the start and end dates for each event, their title and description.

Timeline JS also looks for an image to be displayed alongside the text for each event. Unfortunately, the British Museum’s triplestore does not contain any images that could be used for this purpose. A number of these individuals do, however, appear as Wikipedia entries and these records often include an image of the subject. Given that DBpedia is the Semantic Web mirror of Wikipedia’s contents, it is possible to ask DBpedia for an image URL for each of the 34 contributors using the query shown in Table 37.225 The information returned can then be added to the Google spreadsheet document and the Timeline JS visualisation will update accordingly.

```
1 SELECT DISTINCT * WHERE {
2   ?s foaf:name 'Hormuzd Rassam'@en .
3   ?s foaf:depiction ?depiction
4 }
```

Table 37: retrieving images of people from DBpedia using SPARQL

The cuneiform transliterations and translations

Of course, the primary focus of scholars who have looked at the British Museum’s cuneiform collection over the years has been on the meaning encoded in their script and it would be remiss not to consider to some degree how the cuneiform text might be interrogated using a SPARQL workflow.226 The SPARQL query reproduced in Table 38

---

225 This query was sent to http://dbpedia.org/sparql.
226 This section should be read on the understanding that I am not an expert in the languages written using the cuneiform script. Any observations made and interpretations drawn are derived from a perspective that sees the transliteration and translation texts as arrays of independent word units and not as a series of documents in the sense that a linguist would understand them.
asks the server for the Latin alphabetic transliterations of each of the inscriptions in the collection and the result set includes 2,079 different transliterations.

```
1 SELECT ?s ?transLit WHERE {
2   ?s a ecrm:E34_Inscription .
4   ?s bmo:PX_has_transliteration ?transLit
5 } 
```

Table 38: retrieving the transliterations of the British Museum’s cuneiform inscriptions using SPARQL

The creation of transliterations of cuneiform inscriptions is a complex multi-staged task (Cohen et al. 2004; Proust 2009). It involves the translation of the cuneiform script into another script (often the Latin alphabet) and given the often fragmentary and archaic nature of the source material, it is a process, which involves a huge investment of skilled human labour and domain knowledge. As a result, it is an inherently subjective and interpretive process. For this example, we are interested in the outputs of this process and not the process itself. A sample transliteration, in this case of an inscription found on the brick <http://collection.britishmuseum.org/id/object/WCO92445>, discovered at the site of Eridu in Iraq, is provided below.

(1) {d}amar – {d}suena (2) {d}en-lìl-le (3) nibru{ki}-a (4) mu-pà-da (5) sag-uâ (6) é (d)en- lìla-ka (7) lugal kala-ga (8) lugal urî{ki}-ma (9) lugal an-uba-da lìmû-ba-ka (10) {d}en-ki (11) lugal ki-âg-gà-nì-ir (12) abzu ki-âg-gà-nì (13) mu-na-du

In the case of the Eridu brick, there are thirteen lines of text included in the data returned from the British Museum’s SPARQL endpoint. This transliteration and others like it can then be used to create translations into a modern language text, thereby allowing its meaning to be used to derive valuable insights into particular aspects of ancient practices and lives, whether they be administrative, religious or social.

---

227 The integers contained within the brackets indicate the line number of the transliterated text.
We can look at these units of transliteration in their totality by subjecting them to some basic statistical analysis. The graphic shown in Figure 150 is a visualisation of the frequency of word use within the inscriptions. The most commonly used word is KI (Borger 2003 pt. 737). In Sumerian this means ‘earth’ on its own but has a range of meanings when used in conjunction with other signs. The next most common word, LUGAL, which is made up of LU and GAL meaning ‘big’ and ‘man’, is more straightforward (Jacobsen 2009). In Sumerian, it was used to signify leader or owner in a more general sense and later it came to refer more specifically to a king figure.²²⁸

In certain cases, the RDF cuneiform resources also include English translations of these transliterations, although this information is far patchier. This data is represented as a series of notes, which are attached via the ecrm:P3_has_note predicate to the RDF inscription resource. While there are a large number of these notes (21,818 to be precise), not all of them contain English translations of the transliterations. Most of the notes are given over to other aspects of the inscriptions, such as the genre of the text, observations about the physical characteristics of the artefact and so on. Beyond manually checking each note entry, it is difficult to see how else these data could be sorted into different categories of information.

²²⁸ Of the other most frequently used words, DUMU can mean the number 25 or ‘son’ in Sumerian and ŠAR is a unit of measurement in Sumerian.
Another tactic, and the approach chosen taken by this case study, is to consider the notes as an entire corpus of nearly 180,000 words, as we did for the transliterations, in an effort to discern some broad conceptual patterns. While we were restricted in the types of processing that could be applied to extract meaningful information from the transliteration data, the notes would appear to be far more suitable to more advanced forms of analyses, principally because they are written in English. Natural Language Processing techniques offer a good starting point at which to investigate this potential. Named Entity Recognition is a NLP approach often applied to blocks of text such as this, in order to extract its semantic structure (Ritter et al. 2011). NER essentially parses through a block of text and considers whether each word, on the basis of its context within the overall text and by checking against categorised thesauri, can be assigned a particular conceptual class. For this experiment, the cuneiform notes were run through the Stanford Named Entity Recognizer application (Finkel et al. 2005). This parsed the text and marked up any words which it believed to be one of 3 different classes of elements: PERSON, ORGANIZATION and LOCATION. The results of this processing are shown in Table 39.

<table>
<thead>
<tr>
<th>Class</th>
<th>Value count</th>
<th>Unique value count</th>
</tr>
</thead>
<tbody>
<tr>
<td>PERSON</td>
<td>1,637</td>
<td>481</td>
</tr>
<tr>
<td>ORGANIZATION</td>
<td>1,475</td>
<td>141</td>
</tr>
<tr>
<td>LOCATION</td>
<td>826</td>
<td>365</td>
</tr>
</tbody>
</table>

Table 39: results of running the cuneiform notes through the Stanford Named Entity Recognizer tool

The Stanford NER performed reasonably well when asked to retrieve PERSON and LOCATION word classes from the cuneiform notes but did less well when identifying...
words of the ORGANIZATION word class. The relative frequencies of these PERSON (Figure 151) and LOCATION (Figure 152) words are visualised here as two separate Wordle word clouds. While there are clear instances where the Stanford algorithm has made mistakes, for the most part the words that it has identified with most frequency are as described. For example, Nebuchadnezzar occurs 96 times in the notes and its identification as a PERSON class is correct. On the other hand, the word ‘tablet’, ‘short’ and ‘column’ also appear in this list and these are obvious errors that would need to be filtered out. The words classed as LOCATION are also not without their problematic entries. For instance, the word ‘Nabu’ is included in this list, when in fact Nabu is the name of the Mesopotamian scribe god, later to become the Assyrian and Babylonian god of wisdom (Bienkowski & Millard 2000 p. 206). As such, Nabu should be included in the PERSON list instead. The same problem is noted in the case of Hammurapi and Gudea.

Figure 152: a Wordle visualisation showing the relative frequencies of words in the cuneiform notes identified by the Stanford NER application as being of the LOCATION class

Despite these errors, the rough classification created by the Stanford NER application does serve to highlight certain patterns within the texts. For example, Nebuchadnezzar is
by far the most dominant personality to appear in them and Egypt and Babylon situate the texts more so than any other location. Perhaps, these insights might be suspected by the scholar before they started their analysis but there is a meaningful difference between intuiting that a hypothesis is correct and testing that hypothesis using real data. And in that sense, these results have real value for the scholar of the ancient world. If nothing else, they mark a starting point at which the corpus can be approached in a more systematic way and in many cases this should be the desired outcome of the application of analytical approaches to Archaeological Semantic Web resources.

Research findings

This case study has produced a varied range of research outcomes. To begin, there is the interactive map that shows the find site distributions for all of the cuneiform tablets found within the British Museum’s collection. This map highlights the sites (Nineveh and Sippar), which have contributed the majority of the collection’s artefacts. The case study also yielded a list of the key individuals associated with the delivery of these artefact groups. Following the methodological trend seen in all of the case studies presented in this chapter, these individuals were visualised using an interactive timeline. Using contextual information gleaned from within the British Museum and from DBpedia, this timeline presents a succinct narrative of the museum’s cuneiform collection biography. The case study was also able to present some basic named entity statistics for the tablet’s cuneiform inscriptions.

Any Assyriologist reviewing these findings will no doubt comment that they represent a series of high-level summations of the museum’s cuneiform collection. While I would largely agree with this contention, I would argue that supposition is very different from quantitative verification and the point should not be that these findings are obvious but that the British Museum’s exposure of these digital resources through a SPARQL interface has allowed these facts to be easily verified. Obviously, we no longer live in a pre-digital world and to compare the SPARQL workflow with the gathering of this type of information using traditional archive-mining techniques alone would be of limited use. But it is fair to compare and contrast the SPARQL method to that of the user interface employed by other humanities digital archives. I believe that this case study has shown that the SPARQL methodology elevates the interrogative freedom of the digital researcher far beyond that offered by the alternatives.
Finally, while this case study was solely concerned with an analysis of the British Museum’s cuneiform tablet collection as it is presented via the institution’s SPARQL endpoint, the techniques employed could easily be applied to the study of other material groups housed by the museum. For instance, the British Museum holds an impressive array of ancient ceramics from the Eastern Mediterranean region. These could be approached from a variety of different research perspectives (typological, chronological, geospatial, historiographical).

**Conclusions**

The accounts of the small data mining projects that appear in this chapter show, if nothing else, that consuming Archaeological Semantic Web resources is an iterative process. It involves a lot of to and fro between the questioner, the method and the source material. In this context, data mining of this sort can be viewed as largely analogous to the traditional method of engaging with an academic subject matter (Creswell 2013; Sue Stone 1982). Another insight to be taken from this review is that accessing these resources does not come without its costs. All of these projects demanded that the researcher be familiar and possibly even an expert in the application of a varied number of techniques.

The researcher cannot begin the questioning process without access to the data and so the first requirement of the data mining process is that they find the material that is most suited to answering their question. It is clear that the Archaeological Semantic Web has some way to go before it can be approached as one monolithic queryable dataset as the document web is now accessible. For instance, there is currently no technical way that a SPARQL query can be addressed to all of the datasets that are considered to be part of the Archaeological Semantic Web. This means that the person who poses the question must move iteratively from one online resource to the next as their question evolves. We saw this in the British Museum example, as data is first mined using the museum’s RDF graph. This value was obtained by sending the following SPARQL query to the British Museum’s SPARQL endpoint:

```sparql
SELECT (COUNT(DISTINCT ?artefact) AS ?artefactCount) WHERE {
?artefact ecrm:P108i_was_produced_by ?prodEvent .
?prod ecrm:P7_took_place_at ?place .
?place skos:broader ?placeBroader .
?placeBroader skos:prefLabel "Greece"
}
```

229 As of this date, 14,460 Greek artefacts are represented as resources within the British Museum’s RDF graph. This value was obtained by sending the following SPARQL query to the British Museum’s SPARQL endpoint:
SPARQL endpoint and then it is supplemented with additional images found on DBpedia. The fact that one query interface cannot currently be used to access all of these sources of data, is seriously limiting the Archaeological Semantic Web’s potential as a humanities research tool, and this is a subject that we will return to in Chapter 8.

Secondly, while most Archaeological Semantic Web resources can be accessed in more than one way, SPARQL is by far the most flexible of these approaches. The SPARQL query language is, however, a challenging skillset to master, demanding as it does significant investment in training and practice. The construction of meaningful SPARQL queries also demands a certain degree of domain knowledge on the part of the researcher. In simple terms, the more prior knowledge that a client has about a dataset, the more value they will able to access from that resource. The fact that the British Museum dataset was structured using the CIDOC CRM reduced the amount of time needed before we could access data relevant to our question. This example also showed us, however, that supposed adherence to a public ontology does not always guarantee that the implementing dataset will be fully compliant with that model. As such, data mining can also be seen as something of an exercise in puzzle solving and the speed by which each new challenge is overcome increases as the researcher gains experience by engaging with more Archaeological Semantic Web resources.

Even if all datasets complied perfectly with a public ontology, this would not remove every challenge that can be encountered during the learning of a new dataset’s structure. For one thing, it would be simply impossible to imagine that any one researcher could know the workings of all Archaeological Semantic Web models. Thankfully, ontologies provide a mechanism whereby even the structures of resources, which are at first unknown to the researcher, can be unveiled. As was discussed in Chapter 3, ontologies themselves are published and described as RDF following the OWL or RDFS schemas. Therefore, as long as a client knows how to interpret these ‘meta’ ontologies, they will be able to learn the structure of the ontology that they model. In summary, being able to uncover the structure and meaning of a previously unknown dataset through its published ontology is an absolute requirement, if a researcher is to maximise their data mining exercise.
This chapter also suggests that data mining and data visualisation are two sides of the one interpretive coin (Hyman & Renn 2012 p. 16). It is difficult to perform one set of processes without also being constantly aware of the implications that these actions have on the other. Questions are posed and data is derived from these queries. This data is inputted into a data visualisation or a set of visualisations, which inevitably produce a new set of questions and the hermeneutic circle begins again (Heller 1989). They can help reinforce the hypotheses of a researcher but often as John Tukey, the American mathematician reminds us, “the greatest value of a picture is when it forces us to notice what we never expected to see” (Tukey 1977 p. 594).

From a functionalist point of view, the increasing importance of data visualisation in the praxis of the working digital humanist (Berry 2011 pp. 12–14) demands that this group of researchers be familiar with the possibilities afforded by a whole range of new technologies. They also need to be able to quickly obtain the skills needed to employ these technologies as and when the opportunity demands it. In the three case studies that appeared in this chapter, a myriad range of visualisation techniques was employed, each chosen on the basis of the character of the data to hand and the question being posed. For example, the British Museum case study used the Wordle visualisation tool to highlight patterns within its textual data. While at other times, the representation of the time-based data was accomplished using the Timeline JS tool, which in turn demanded knowledge in a whole different set of skillsets. The Priniatikos Pyrgos loomweight spatial analysis and Trench 2 stratigraphy case studies involved a different set of visualisation knowledge again.

All of these methods took time and effort to master and it would be disingenuous to claim that this investment was negligible. At the same time, however, it would be difficult to argue that the approach taken by these case studies was in any way superfluous or indulgent to the needs of their questions. The results of these analyses required that they be presented not just as end products but also as devices that would aid and direct interpretation right throughout the research process (Berggren & Hodder 2003).

In summary, the creation of new knowledge is a complex matter involving technological, methodological, epistemological and sociological variables. The question that emerges from this chapter is whether the employment of Archaeological Semantic Web
researcher methods has any implications when viewed through these various lenses of enquiry. And if it does, what do these implications mean for the way that we construct, disseminate and internalise narratives about the past? It is to this subject that we turn in the next chapter.
Chapter 7 – A critical analysis of the Archaeological Semantic Web

Considering the sociological and epistemological implications of the Archaeological Semantic Web

‘Over the past few years I've had an uncomfortable sense that someone, or something, has been tinkering with my brain, remapping the neural circuitry, reprogramming the memory. My mind isn’t going—so far as I can tell—but it’s changing.’
(Carr 2008 p. 89)

‘We need never be hypnotized by the computer's capacity to count into thinking that once we have counted things we understand them.’
(Computers and the Humanities 1966 p. 1)

Introduction

In many technology-driven narratives the question of ‘why’ can come to be framed within the language of technical expediency and functional utilitarianism (Evans & Rees 2012 p. 25). The why becomes an attempt to link a functional need to an applied method. And while these approaches provide value in answering certain types of questions, what they fail to understand is that technology is fundamentally a human-centred activity (Berry 2012 p. 17; Frabetti 2012). While the heavy computational lifting may be done by a machine, the processes that constitute this machine are directed by human instruction, and they work together towards achieving an outcome that will be ultimately meaningful to a human (Hayles 2012 p. 47). In the preceding six chapters of this work, we have introduced the Archaeological Semantic Web as ‘the’ solution to the problem of heterogeneous digital data within archaeology. And while an attempt has been made to foreground the consumer in this knowledge system, even here the tendency has been to present the wants, needs and results of using Semantic Web techniques from this functionalist perspective. The temptation to treat the consumer of the Archaeological Semantic Web as a generic being with universal needs and expectations, who responds to outcomes in particular predetermined ways is as strong as it is unhelpful. As such, this chapter is an attempt to correct this functionalist bias by considering what it means from a sociological as well as an epistemological point of view for an archaeologist to use Semantic Web technologies.
We begin by comparing the Archaeological Semantic Web method with other systems that might conceivably be employed to deliver a comparable outcome. Next, we present the Archaeological Semantic Web from an explicitly sociological viewpoint. We begin by reviewing existing sociological research relevant to our study. This is followed by an account of a series of data gathering experiments carried out specifically for this project. These surveys and workshops were designed to assess the archaeological and cultural heritage community’s opinions on data sharing, data collaboration and Archaeological Semantic Web use. The emphasis in this section is on the inspection of the Archaeological Semantic Web as it actually is and not as it is envisioned to be. We conclude this discussion with the presentation of a model for the use of the Archaeological Semantic Web, which distils the findings of the previous sociological studies.

The second half of the chapter is concerned with the Archaeological Semantic Web as a living and breathing knowledge system. It asks how knowledge produced by the Archaeological Semantic Web can be and is differentiated from pre-Semantic Web and pre-digital archaeological knowledge (Rieder & Röhle 2012). Ultimately, this concluding section asks how our Archaeological Semantic Web use model fits within an existing archaeological epistemological system and what significance, if any, does adherence to this system entail for archaeological schools of philosophical thought in general.

**Comparing the Semantic Web to the alternatives**

The Semantic Web, as has been reiterated throughout this work, is a complex animal involving the intersection and collaboration of a number of different processes (Figure 153). To start, there is the matter of how Semantic Web data is sourced. We call this the Knowledge Acquisition phase and whereas this was once carried out predominantly using human labour, now we are increasingly seeing this aspect of the Semantic Web’s workflow being handled by machines using techniques such as Natural Language Processing and image processing (Gaines 2013 p. 146). Next, the Knowledge Representation process takes over by ensuring that the information gathered is presented in a form that a machine can understand (Sowa 1999 p. xi). It is here that we see the employment of RDF and higher-order ontologies to give different levels of structure to the dataset. Knowledge Management takes this data and ensures that it is stored and maintained so that it can be sustained over time and as it grows in scale (Benjamins et al.
1998 p. 5.1). Of course, data is of little use if it is never accessed. The final process, Knowledge Retrieval, enables users (machine or human) to ask questions of a dataset using a combination of reasoning techniques.

Each of these phases is sequential and performs functionally separate tasks. On the other hand, each phase is dependent on the output of the previous phase. Without Knowledge Representation, there is nothing for Knowledge Management to manage and output on to the Knowledge Retrieval phase and so on. In all of these various categories of activity, the human designer is presented with choices, which in turn impact on the options available to the data consumer. For instance, it is not a requirement that Semantic Web data be represented using the RDF format or that OWL be used to author its ontologies. These options are adopted as others are discarded. Essentially, what this means is that the Semantic Web is not a standard in which all of its constituent parts need to be adopted wholesale. The Semantic Web implementer is free to pick and choose those aspects of the model that suit their need.

In this section, we look at some of these alternatives, which are not generally considered to be compliant with the Semantic Web ethos. While it would be convenient to place all of these various alternatives in binary opposition to the techniques outlined in this thesis, we will see that often no clear-cut distinction can be drawn between the two.
FileMaker Pro and offline Archaeological Information Systems

The FileMaker suite of products has been an immensely popular information system choice of the archaeological community since it first appeared on the market in 1983 (Solid IT GMBH 2015). Now in its 14th major version release, the system is available as a server, client and iOS app for mobile devices. FileMaker is an RDBMS, which means that each of its databases is divided into separate tables, and each table contains a set of records that can relate to the records in other tables. The system boasts an impressive feature list, including multi-language support, scripting and the provision of SQL and ODBC interfaces, which allow non-FileMaker clients to consume the data that it hosts.

While many RDBMS database desktop applications, such as Access, represent their table’s two-dimensional content as an abstract grid of rows and columns, FileMaker since the 1980s has been allowing its data to be represented using a graphical interface (Figure 154). This interface can be designed by the creator of the database (Prosser & Gripman 2010 pp. 97–126).

While the elements of a FileMaker table design – text fields, list boxes, radio boxes, image containers, et cetera – are seen in most FileMaker databases, it is their unique arrangement in ways that suit the specific application in mind, that has made the system so popular among archaeologists. A browse through the paper submissions to any of the
CAA conferences held over the last decade shows how popular FileMaker has become within this community. Famously, in 2010 Apple showcased the work of the University of Cincinnati’s team of excavators at the Roman town of Pompeii (Apple Inc. 2010; Ellis & Wallrodt 2011a, 2011b). Ellis and his team employed the use of the FileMaker Go app running on the first iPad model to document their excavation of a number of insulae at the site. Most interestingly from a field recording perspective, the project members used this setup to draw their sections, elevations and plans, as well. This was one of the first attested uses of a digital tablet device in the making of archaeological line survey drawings.

The basic design of FileMaker was developed in the pre-Semantic Web era and there has been no noticeable effort in the intervening years to move its design philosophy more towards that of the Semantic Web. Since 1994, it has been possible to publish data created using the system onto the web using a web server that is bundled with the product. However, making data available through a proprietary web interface does not necessarily make it part of the Semantic Web. Were one to rate a notional FileMaker database using Berners-Lee’s 5-Star system, it could be awarded one star, if it was published online under an open access and use license. The data contained within a FileMaker database is also structured and so it would be eligible for a second star. However, the format of this structure does not follow an open standard and so it could not be given a third star. It is, however, possible to use URIs to denote data resources within a FileMaker database but this addressing cannot be used to allow other data to link into the FileMaker database from outside. In the same way, it would be impossible to create links between the FileMaker records and external Semantic Web resources.

FileMaker also allows for collaborative working, which means that more than one user at a time can access and work on a single networked database. In an archaeological context, this can be a big advantage when dealing with the manual digitisation of paper records, which has become such a feature of the archaeological Knowledge Acquisition process. However, despite these positives, FileMaker suffers from the fact that it is a commercial product and, therefore, comes with a commercial price tag, which for many archaeological projects represents an insurmountable obstacle, particularly for those that need to manage the digitization of large quantities of data, which would necessitate the

230 See Arleyn et al. (2008), Kondo et al. (2012), Motz and Carrier (2013) for examples.
purchase of a number of unit licenses.\textsuperscript{231} There is also the fact that FileMaker is a closed proprietary system. Its code is kept under the control of Apple, who has owned the product since the late 1980s. This means that it is impossible to extend the product’s baseline feature-set by adding custom functions, which might be demanded by a specific application.\textsuperscript{232} The formats of the data are also proprietary, although it is possible to export portions of this information as open formats.

Ultimately, FileMaker can more than adequately fulfil the function of an Archaeological Information System in environments in which it is deemed sufficient for the project’s data to be exchanged and shared among a small user group. Despite stated desires to the contrary, this can often be the reality for most archaeological projects that work in often inaccessible and internet-excluded parts of the world. Technological advances over the coming years will, however, increasingly allow more archaeological sites to establish on-site digital data networks. These networks may even have connections to the internet, and in this scenario the FileMaker closed user-group model becomes a much less attractive option.

\textbf{MySQL and the closed online Archaeological Information System}

If you have ever visited a web page, then the chances are that you will have interacted with a MySQL database server. So ubiquitous has MySQL become that it is perhaps better to speak about an RDBMS space that is divided between MySQL and all of the rest.\textsuperscript{233} The MySQL system was first released by a Swedish company, MySQL AB in 1995. Sun Microsystems bought the company in 2008 and two years later Oracle, MySQL’s current owner, acquired Sun. SQL (pronounced ‘sequel’) stands for Structured Query Language. It is the formal language used to interact with the data contained within

\textsuperscript{231} It may seem somewhat trite to weight financial cost so highly in the evaluation of the FileMaker product. However, it is my experience that price can easily override all other considerations in real-world archaeological field projects scenarios. While this policy often proves to be shortsighted and is the cause of various subsequent problems, it is, nonetheless, a working reality for many archaeological initiatives.

\textsuperscript{232} This inability to extend FileMaker was one of the primary reasons why the Priniatikos Pyrgos project made the decision to move away from a FileMaker data infrastructure and to build the Open-source Software and Linked Open Data linkedarc.net system.

\textsuperscript{233} This analysis excludes SQLite, which is slightly different to other RDBMSs in that it is not a client-server application (Owens & Allen 2010 p. 1). In percentage terms, SQLite is actually the most widely used RDBMS in the world today, as it is embedded into every Chrome and Firefox installation. ScaleBase recently estimated that MySQL held 56\% of the world’s database market share (Campaniello 2014). And MariaDB, which was next in the list at 18\%, is an offshoot of MySQL and is currently managed by one of MySQL’s founding developers, Michael Widenius (Bartholomew 2013).
certain RDBMSs, and its syntax is very similar to that of SPARQL. An example of a simple SQL statement is shown in Table 40.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SELECT *</td>
</tr>
<tr>
<td>2</td>
<td>FROM Book</td>
</tr>
<tr>
<td>3</td>
<td>WHERE price &gt; 100.00</td>
</tr>
<tr>
<td>4</td>
<td>ORDER BY title;</td>
</tr>
</tbody>
</table>

Table 40: a simple SQL statement example

It is difficult to identify a single reason why MySQL, and not one of the many other RDBMSs,\(^{234}\) has come to dominate the database market, particularly the web server market. However, it is likely that timing played a large part in its success. When MySQL first appeared in 1995, the web was still in its infancy and searching for technological stacks with which to build the basic functionality that we now take for granted of the modern web site. While the web model, as proposed by Berners-Lee and Cailliau, described a document web in which the contents would be made up of documents stored as files on a web server’s file system, it soon became apparent that a more flexible and data-centric storage architecture would also be required.

Of course, this talk of data in contrast to documents is a familiar concept in this thesis, as it is one of the primary drivers of the Semantic Web. Having said that, the data-need addressed by the invention of MySQL is subtly different to that which Berners-Lee and Cailliau hoped the Semantic Web would meet. MySQL proved hugely successful as a means of delivering data storage and access facilities to an emerging online data-driven website space. This was particularly the case for those sites, which were interested in exploiting the web as an e-commerce platform (Tian & Stewart 2006 p. 560).

While MySQL is like the RDF-based triplestores of the Semantic Web, in that it stores information at a fine level of detail, it diverges from the Linked Open Data model by hiding its contents from the outside world. A MySQL database has one entry-point into its contents and that is its SQL interface (Figure 155). It will accept no other form of engagement. This ‘hard shell’ database model has given rise to the view of the MySQL database as a data silo and the web as an archipelago of isolated MySQL databases, all ignorant of each other’s contents (Cattell 2011 p. 13). Whereas the Semantic Web dataset is characterised by its interconnectedness, its promotion of data re-use, its dynamism and

---

\(^{234}\) For instance, Firebird, IBM’s DB2, Apache Derby or PostgreSQL.
Despite these limitations, the MySQL solution has been used extensively, often as part of the so-called LAMP (Linux, Apache, MySQL, Perl/PHP/Python) configuration (Lee & Ware 2002), as the information store in a wide range of Archaeological Information System designs (Eve & Hunt 2008; Hochin et al. 2009). Archaeologists are attracted to the LAMP configuration primarily because it is low or zero cost, stable, well documented and capable of handling large amounts of data. Also, the RDBMS model, as a way of conceptualising data, is familiar to and compatible with the praxis of the modern archaeologist. Archaeologists have been categorising information into associated groups based on their material characteristics since the discipline’s inception (Rouse 1960), and this practice sits well with the typological demarcation that MySQL’s table design...
promotes. This familiarity allows archaeologists to very quickly translate internal conceptualisations representing archaeological material realities into MySQL data structures.

As far as the Berners-Lee 5-Stars of Open Data rating goes, a MySQL database fairs slightly better than was the case for FileMaker. As with the latter, MySQL databases can be easily published online under an open license and this data is also inherently structured. Unlike FileMaker databases, open formats can be used to structure its data, which gives it another star. It can also use URIs to identify its records but again URIs cannot be used to provide links between the database’s content and external data.

Another consequence of building an Archaeological Information System on top of MySQL is that the database’s ontology must be decided upon and fixed at the point of the database’s creation. As we have seen throughout this work, the archaeological process produces data of a wide variety of types. There are the conceptual units (contexts, periods, types) that represent the excavation, stratigraphic analysis and interpretive processes, and then there are those concepts, which map more directly onto the tangible products (samples, artefacts and empirical readings) of excavation and survey. Each of these concepts would need to be modelled as a separate table in a MySQL database, with each instance of the concept being connected to instances in

Figure 156: an example MySQL archaeological database
other tables using the records’ unique identifiers, as is shown in the example database structure in Figure 156. This works well in a scenario in which the database’s ontology can remain unchanged throughout its life. From the user’s perspective, the writing of queries that ask for information from various different tables is a relatively straightforward matter. This is shown in an example (Table 41), which asks our hypothetical MySQL database containing four tables (contexts, artefacts, types and periods) for a description of all of the Bronze Age daggers found in Context 1.

```sql
SELECT t_artifact.t_artifact_description, t_type.t_type_name, t_context.t_context_name, t_period.t_period_name
FROM t_artifact
INNER JOIN t_type
ON t_type.t_type_id = t_artifact.fk_t_type_id
INNER JOIN t_context
ON t_context.t_context_id = t_artifact.fk_t_context_id
INNER JOIN t_period
ON t_period.t_period_id = t_artifact.fk_t_period_id
WHERE t_period.t_period_name = 'bronze age'
AND t_type.t_type_name = 'dagger'
AND t_context.t_context_name = 'context 1'
;
```

Table 41: posing a multi-table query to a SQL database

While the need to include the joining logic makes the query slightly more convoluted than would be the case had it been asked of an RDF dataset and written in SPARQL, the sense of the statement is not overly opaque. However, the success of MySQL in modelling static ontologies breaks down once you start to deal with the more fluid ontological structures that are a feature of the Semantic Web. For instance, a Semantic Web resource that represents Context 1 in the example shown above might need to be expanded upon to include links to geo-coordinate data as it becomes available at a later point in the archaeological project’s work. To do this using RDF triples requires a small amount of effort on the part of the data creator. The Context 1 URI need only be linked to a new URI, which represents the new geo-coordinate data. To achieve the same result using a MySQL database would require a change to the structure of the table that stores
the context data and perhaps the addition of a new table to house the geo-coordinate information. While affecting this sort of change in the MySQL database’s structure or schema might sound quite trivial, in reality it will most likely demand that a chain of related modifications also be made. These alterations take time and demand the expenditure of resources, which may or may not be possible in the circumstances.

**NoSQL and Big Data Archaeological Information Systems**

The NoSQL database is the newest of the three information system designs included in this review. It is also most frequently associated with the management of Big Data. As was discussed in Chapter 2, Big Data defines data that is, among other things, largely unstructured. This lack of structure is something that is increasingly seen to characterise the majority of the data that we find on the web and because of that both Big Data and NoSQL are attracting quite a bit of attention recently (Abramova et al. 2014; Mayer-Schönberger & Cukier 2013 p. 45). While Archaeological Information Systems based on the NoSQL type are still something of a rarity at this early stage in their development, it has been suggested that archaeological data’s inherent messiness and its propensity to employ informal ontologies ensures that it will become an obvious candidate for management by NoSQL systems in the coming years (Gattiglia 2015 p. 5).

With any innovation, there are, however, always early adopters and the OpenDig Archaeological Information System is one such pioneering example (Vincent et al. 2014). OpenDig uses the Apache CouchDB NoSQL system (Anderson et al. 2010) to deliver both excavation and post-excavation digital recording and analysis tools to a distributed team of University of California archaeologists working on the Madaba Plains Project, which is conducting a series of field excavations in the area to the east of the Dead Sea in Jordan (Herr 1993).

NoSQL databases use document-based architectures. Each document in these systems is allowed to conform to its own data schema or ontology. This ontological fluidity means that they are ideally placed to manage the hybridity of an increasingly multidisciplinary archaeological practice (Greene & Moore 2010 p. 192; Tite 1991). However, while it would be tempting to see NoSQL systems as Big Data equivalents of RDF triplestores because of this liberal attitude towards data ontologies, the comparison is not entirely appropriate. Semantic Web data is, if nothing else, highly structured despite its ability to change that structure as and when adaption is needed. Big Data, on the other hand, is defined by data variety and by its ability to manage this messiness in the best way.
possible (Gattiglia 2015 p. 5). In that context, it is likely that NoSQL and RDF solutions will co-exist with one another for the foreseeable future, each servicing a particular need of the online digital data network.

Assessing a NoSQL system against Berners-Lee’s 5 stars of Open Data criteria is a more subjective matter than was the case for MySQL and FileMaker. This is principally because NoSQL is still very much an evolving technology and, as such, its philosophical bases have yet to be clearly defined. In principle, it is possible to see a NoSQL system being awarded 5 Open Data stars but the problem is that there is no level of standard practice within the NoSQL sector. While efforts are under way to correct this (Clark 2014), there is currently no such thing as a typical NoSQL system, as there is a typical RDF triplestore. Because of this, NoSQL systems are currently proprietary solutions and, as such, they can be awarded no more than 2 Open Data stars.

**Alternatives in review**

Comparing these three information systems against Berners-Lee’s 5 Stars of Open Data (Figure 157) is perhaps somewhat unfair. Each of these alternatives to the Semantic Web has been developed within its own social and technological contexts. They service their own particular needs, which are not necessarily relevant to the objectives of the Semantic Web. It is perhaps better to ask how these systems might be employed in answering the central question posed by this thesis. When viewed from this perspective, FileMaker presents a number of difficulties. First, because of its pricing structure, it is not conceivable to imagine that it might ever attract the sort of wide-scale usage that would
be necessary, if it were to be the information system with which all archaeological digital data was to be managed, exchanged and analysed. The proprietary or closed nature of FileMaker also means that it will also present problems as part of a multi-solution system. MySQL does not present the same problems of cost and its lengthy development history has produced a well-known and well-built design architecture that is more than capable of dealing with archaeological data. The problem with MySQL is that it works substantially less well in environments where change and hybridity are the norm, which is exactly the type of data environment that this thesis’s enquiry is premised upon. There is also the point that archaeological data stored in the RDBMS’s rigid form becomes an unwieldy behemoth of multiple tables each containing hundreds of columns, most of which are left empty and irrelevant for the vast majority of records. And ultimately, the fact that this data can only be accessed by a single SQL interface point of access severely limits its usefulness as part of an Open Data infrastructure. For its part, it is simply too early to say how NoSQL databases will affect the Archaeological Information System sector and what their impact, positive or otherwise, will be on the matter of heterogeneous archaeological data.

It is far more likely that no one information system will come to dominate the archaeological data ecosystem any time soon and that instead we will see a continuation of the current state of affairs, in which all of the information systems mentioned above as well as new entrants will compete with one another for market share. To accept this statement, does not undermine the basic position taken by this thesis that the Archaeological Semantic Web presents the best way forward for Archaeological Information Systems. However, it does certainly concede that the Archaeological Semantic Web will always operate within a multi-solution environment.

**A sociological perspective of the Archaeological Semantic Web**

In attempting to find a solution to this problem of heterogeneous archaeological data, it has become increasingly apparent that whatever option is selected must also take into account the relevant sociological factors. Of course, the basic idea that a technical challenge is affected by non-technical influences is not a new idea. It has long been argued that society, politics, economy, technology, religion and any other category of anthropocentric activity are involved in a constant conversation and exchange of ideas and influences (Bijker et al. 2012 p. xlii). And this point has been made by a number of commentators about the Semantic Web in broad terms (Gruber 2008; Hendler et al.
2008 p. 68) and about the Archaeological Semantic Web more specifically (Isaksen 2011 p. 13). This following section presents an overview of the latest sociological research relevant to the Archaeological Semantic Web. It also recounts the findings of my own research on the matter, which took the form of an online survey addressed at the archaeological and cultural heritage communities as well as a series of targeted workshops.

**Conclusions**

- The ARIADNE project is highly relevant in terms of addressing important user needs.
- The research community longs for better transparency of available data and, equally, for improvements in data accessibility.
- Data and metadata quality are also relevant concerns for researchers and, in particular, for data managers.
- The complex diversity and fragmentation of institutional “data habitats” will be a major challenge for the project.

**Specific conclusions (see 8.1.2 / 8.1.3)**

- New tools for humanities researchers, to be accepted, should have a low learning curve (ease-of-use) and offer immediate efficiency gains in their existing routines.
- There are two major barriers for sharing data with other researchers in a repository: a perceived lack of recognition for sharing, and the (additional) work effort for preparing the data set so that it can be deposited.
- When searching data, it is not so much the source as such that matters – it is the quality of the data contained.
- Only few respondents feel that the current level of online availability of research data is satisfactory.

**Expectations towards ARIADNE (see 8.1.3)**

- The central expectation of the user communities (researchers, data managers) is that ARIADNE should provide a better overview of existing data resources.
- Search-portal functionalities (that facilitate this overview) were the top-rated services, in particular by researchers, which ARIADNE could deliver.

**Conclusions on specific data types by SIGs (see 8.1.4)**

- Grey literature: promote guidelines for digitisation, define a core metadata standard for grey lift, conduct R&D on novel indexing and extraction technology.
- Excavation and monuments data: develop tools and guidance based on international standards, include the effort for long-term curation already in project plans.
- Visual media: Improve interaction with high-resolution images, documentation of complex 3D models, and licensing conditions.

Figure 158: outcomes of the ARIADNE D2.1 report on user needs (ARIADNE 2015a p. 155)

---

235 Note that these studies have been selected on the basis of their explicit relevance to the subject of the user of Archaeological Information Systems. While other humanities fields and the digital humanities have conducted similar research into this topic of user expectations and requirements (Runter & Schonfeld 2012; Stone 1982; Warwick et al. 2007, 2012, 2012), these will not be covered by this review, as it is the position of this thesis that the archaeological researcher exhibits a unique set of practices.
Existing scholarship on the social Archaeological Semantic Web

The ARIADNE project, which was introduced in the context of the CRMarchaeo data model in Chapter 3, has published a number of reports since it was first established and two of these reports are of particular interest to this subject of the Archaeological Semantic Web as a social entity. Its D2.1 “First report on users' needs” (2014) and follow-up D2.2 “Second report on users' needs” (2015a) reports present a contemporary view onto the archaeological user’s experience of producing and consuming digital resources. The intention of both reports is to construct a picture of exactly what it is that users of archaeological digital resources want and expect from these services, so that this information might then be used to construct a list of criteria with which to direct ARIADNE’s building of a Europe-wide infrastructure to integrate each member-state’s archaeological digital data resources (ARIADNE 2014 p. 5). The source materials used by the two reports are ARIADNE’s own independent surveys and the findings of other related studies (Bulger et al. 2011; Harley et al. 2010; Open Context 2009; Research Information Network 2008; Ross et al. 2013 pp. 111–114). And while not all of the resources that are referenced by the respondents can be classified as specifically related to the Archaeological Semantic Web, as it is defined by this thesis, the types of subjects and issues mentioned are, nonetheless, mostly relevant to our study.

Figure 159: the ‘Russian Doll’ model of archaeological data sharing
Figure 158 lists the main outcomes of the first ARIADNE user needs report. The core message communicated by this summary is that there is a definite desire within the archaeological research community to have better access to transparent, high quality data that accommodates the high levels of complexity inherent in the community’s research activities. The report makes a distinction between the different types of archaeological research user that exist – the individual researcher or project research unit, the research institution, the data centre and integrated data services (ARIADNE 2015a pp. 54–58). While all of these user types are clearly important in understanding the social dynamics of the Archaeological Semantic Web, it is the first type, the researcher, that is of primary interest to our study.

Out of the nearly 900 individuals who responded to the D2.1 survey, over 700 of them were archaeological researchers. The picture painted by the report of the data exchange habits of this cohort can be summarised in the diagram shown in Figure 159. This ‘Russian Doll’-style graph network includes the individuals active in the archaeological space and the various domain levels or shells in which they are situated. At the level of the project, there is a willingness and an ability to share data among colleagues. The report and other commentaries on the subject identify these arrangements as informal data sharing networks and their prevalence is regarded as significant (Acord & Harley 2012 p. 4; ARIADNE 2015a p. 47). Many scholars argue that the majority of digital practices have evolved to simply mimic pre-digital analogues (de Bolla 2014; van Peursen et al. 2010 p. 23; Schnapp & Presner 2009). Given that the sharing of information is a fundamental aspect of the way that humans create new knowledge (Carvalho & Goodyear 2014 p. 15), the fact that informal networks are still the dominant way that archaeologists gain access to their digital data needs should not be too surprising (Kansa & Kansa 2013 p. 89). The reports notes that as one moves further away from this inner sharing nucleus, the propensity to exchange data between agents becomes, however, less pronounced (ARIADNE 2014 p. 48). It decreases first as you enter the institutional zone, which governs the individual project, and then the disciplinary knowledge domain, which encompasses the institutions. When you eventually reach the outer realm of global knowledge, there is very little sharing going on at all.

It is necessary, however, to also consider other possible causes for this behaviour beyond these traditions of praxis. There are a number of factors at play, which collectively agitate
for change to the status quo. To begin with, there are the push dynamics, which drive researchers towards greater sharing of their data. For instance, there is an increasing awareness among humanities policymakers concerning the matter of data sharing (Borgman 2012 p. 1059). This ultimately trickles down into the creation of additional data sharing-specific criteria within project funding applications (Griffiths 2009 p. 52). If data sharing were to become a requirement of academic contracts, this would also represent another sizeable push factor. Although clearly this type of requirement is currently a long way off challenging the dominance that traditional publication currently enjoys in this respect (Borgman 2009 p. 14). On the other hand, there are the factors that increasingly serve to pull archaeologists towards greater sharing of their data. First, garnering a reputation as a data sharer might result in an increased professional standing in one’s research community (Fecher et al. 2015 p. 12). Other reports on the subject indicate that sharing of research data also can result in an increase in citations for the reference work, which is clearly of value to the authors of the work (Piwowar 2010). Lastly, the researcher might be motivated by altruism to contribute something of value back into their host research community.

While these push and pull factors increase the potential of there being a greater level of data sharing between individual researchers within the community, there are also factors, which act as obstacles. Cost is considered by 32% of the respondents in the D2.1 report as being a ‘very important’ barrier to their depositing of data in public data repositories (ARIADNE 2014 p. 106). The amount of extra effort required to actually publish the data and the work needed to prepare its ‘metadata’, which is described by the report as being ‘data about data’ (ARIADNE 2014 p. 53), are also significant factors, with 35% and 37% respectively saying that these represent very important considerations. This last point is worth highlighting, as it directly contradicts a trend within the data manager cohort. 63% of this second user group sees ensuring metadata quality as being very important (ARIADNE 2014 p. 130) and, interestingly, 44% also believe very strongly that their users want the data that they access to have a high metadata quality. This picture resonates with the view contained in the history of the CIDOC CRM, which is recounted in Chapter 2. While the CRM is largely championed by data producers, the ARIADNE reports tell us that it has been met with indifference by archaeological data consumers.
Chapter 7 - A critical analysis of the Archaeological Semantic Web

The linkedarc.net ‘Attitudes to Digital Data Sharing within Archaeology’ survey results

linkedarc.net resource

The questions and results of the ‘Attitudes to Digital Data Sharing within Archaeology’ survey

http://linkedarc.net/surveys/arch-datasharing

As well as reviewing existing research, I also conducted my own independent studies of the social aspects of the Archaeological Semantic Web. The intention of these was to learn about the specific expectations and needs of its typical user. To do this, I created a short anonymous online survey composed of 20 questions, which I hosted on the linkedarc.net website. 236 The central objective of the survey was to answer the question: is there an appetite within the archaeological community for digital resources that conform to the Semantic Web model? The answers to the questions posed in the survey produced what are defined as ‘self-reported metrics’ (Tullis & Albert 2008 p. 123), which are data created by the respondent without the intervention of a third party. 237 In order to retrieve as large a sample size as possible, 238 the survey was advertised on social media channels.

---

236 The questions posed by the survey are reproduced in Appendix B. Respondents could choose to submit their email addresses, if they wished to be included in a draw for a €40 Amazon gift card. These email addresses were excluded from all subsequent publications of the data itself and its analysis.

237 For an argument against the validity of the results obtained using self-reported metrics see (Graves 2013).

238 Achieving an acceptable sample is supposed to reduce the potential for committing alpha (finding non-existent patterns in the data) and beta (failing to find actual patterns in the data) survey errors (Peers 1996). What ‘acceptable’ equates to depends on the total size of the population that you are looking to survey, which can be a difficult to assess given that archaeological participants need not be identified in the usual quantifiers such as censuses. Doug Rocks MacQueen has attempted this for US archaeology, proposing that there were in the order of 11,000 individuals employed in academic, federal and private-sector posts in 2014 (Rocks-MacQueen 2014). In Europe, the figures have been gathered in a more systematic way with the Discovering the Archaeologists of Europe 2012-14 project coordinating matters. It was estimated by the body that approximately 33,000 archaeologists worked in the sector across the 21 participating countries in 2014 (Aitchison et al. 2014 p. 6), with the UK representing the largest contingent of just under 6,000 in 2013 (Aitchison & Rocks-Macqueen 2013 p. 10). It is more difficult to get data about the status of the field within other parts of the world. A report on Japanese Archaeology from 2003 estimated that there were about 7,000 ‘specialists’ active in the field (Keally 2003), and a 2009 study of Australian archaeologists estimated that there were at least 399 professional archaeologists working in that country (Ulm et al. 2013 p. 35). As it happened, the majority of the responses came from the US and Europe with smaller numbers coming from the Middle East and Australia. Going on this basis, it seems fair to use these figures.
(Twitter and Facebook) and on various content-relevant mailing lists (Digital Classicist List, Antiquist).

![Figure 160: the location of the ‘Attitudes to Digital Data Sharing within Archaeology’ survey respondents](image)

In total, the survey was answered by 246 unique respondents, who for the most part were located in the US, Europe, the Middle East or Australia (Figure 160). The vast majority (76%) of the sample group worked within the field of either contract or research archaeology, which was the intended demographic (Figure 161). The survey contained two main question groupings. The first of these groupings looked at the level of respondent familiarity with the Semantic Web and related terminologies, and the second set of questions asked the respondents about their expectations of digital data sharing and their own willingness to consume and publish archaeological Open Data.

(excluding the Middle East for which reliable figures could not be found) and adding these together produces a total world population of approximately 45,000 individuals. For surveys that retrieve categorical data, Kotrlik and Higgins (2001 p. 48) suggests that when dealing with a target population of 10,000, the survey maker should try to get a sample of between 264 and 623 depending on the error tolerance required. While not ideal, the sample size of 246 valid responses, which was actually returned, can still, however, provide meaningful results (Tullis & Albert 2008 p. 17).
Chapter 7 - A critical analysis of the Archaeological Semantic Web

Figure 161: ‘Attitudes to Digital Data Sharing within Archaeology’ survey demographics

Figure 162: ‘Attitudes to Digital Data Sharing within Archaeology’ relevance and familiarity
On the familiarity front, the results tell an interesting but perhaps not overly surprising story (Figure 162). Familiarity with the term ‘Big Data’ showed a fairly even spread across the first four options with a noticeable jump in numbers (31.3%) citing a high level of familiarity with the expression. The Open Access responses, for their part, indicated that the term is fairly well embedded within the consciousness of the sample group but this trend is much less pronounced for Open Data. Lastly, the Semantic Web is not at all well understood by the sample, with a dominant 37.3% stating that they were very unfamiliar with its meaning and only 10.4% saying that they were very familiar with it.

While one might expect that there would be a certain degree of consistency between the relevance to work practice figures and the related familiarity figures, this transpired not always to be the case. Open Access does appear to show that familiarity within the minds of the sample can be linked to one’s consuming and/or producing of more Open Access resources. However, the Open Data relevance/familiarity figures show that Open Data’s relevance to the respondents’ work practices surpasses the awareness that the group have of the term. This is an interesting result, as it might suggest that the respondents are expected to use Open Data techniques in their work practice but that they feel that they do not have adequate knowledge of the approach to satisfy this need. The Semantic Web relevance to work practice figure, which showed 41.4% of the sample reporting a neutral stance, confirms that the Semantic Web appears to be irrelevant to the majority of the respondent’s work practices. This result is also telling.

Moving on to the final set of questions (Figure 163), an overwhelming 73.9% of respondents stated that they believed digital data sharing to be very important to the future success of archaeology and 75.9% held that there was significant demand within the community to make this happen. As regards consumption of this data, 91.2% stated that they were enthusiastic or very enthusiastic to have greater access to other archaeological research datasets. This sentiment is reciprocated in the 84.7% of the sample who stated that they are more positive than negative about the principle of making their own data available to the community. This last figure is generally borne out by the declaration by 73.1% of the sample that they had previously shared their own data with the community.
The responses relating to the obstacles that impede the greater sharing of archaeological digital data are also interesting. 90.3% of the sample believes that work practices and/or regulations have a neutral or negative effect on the sharing of data. A strong cohort of 71.9% of the sample believed that lack of digital skills had a significant or very significant negative impact on sharing. However, for the most part, the sample does not appear to link legal/regulatory obligations to data sharing problems.\footnote{One respondent included the following informative comments on the matter of whether regulatory and legal factors have an inhibiting effect on the sharing of data. ‘This really depends on the country you work in. In Greece, the national archaeological authorities are extremely severe about the sharing of any kind of data, even photos, before archaeological publication. In other places, local colleagues will insist that data can’t be shared even if the laws permit it. In still other places, local colleagues will instruct you to go ahead and share data even if they’re not sure whether there’s a law that forbids or permits it … In Italy, the state now mandates submission of reports to Fasti, so there is at least limited data sharing required by law. It’s therefore really hard for me to say “yes” or “no” to this question, because depending on what country I happen to be working in, regulations may either inhibit or facilitate.’}
Finding patterns in the ARIADNE and linkedarc.net survey results

So how do the linkedarc.net survey results tally with the ARIADNE findings, and when viewed in their totality, do they tell us anything new about the sorts of needs, aspirations and challenges that define the archaeological digital data-sharing environment of 2015? The conclusions of both surveys largely agree. There is certainly an appetite at a grassroots level within archaeology to share more data and to gain access to more data (however vague that particular set of aspirations might be – to what end?). While the ARIADNE user needs surveys do not specifically question their sample about Semantic Web related methodologies, there are nonetheless relevant interpretations to be read from the results. To start, data transparency is considered to be important by archaeological researchers but at the same time there is a grassroots rejection of the dogmatic standards-based approach represented by projects such as the CIDOC CRM. This pattern is echoed by the linkedarc.net survey results, in which awareness of the Semantic Web is very low, while, at the same time, the principles that surround the Semantic Web (openness, sharing, re-use) are generally endorsed.

This finding clearly presents a challenge to groups such as ARIADNE who are clearly well disposed to Linked Open Data, public ontologies and vocabularies, all of which are consistent with Archaeological Semantic Web practice (ARIADNE 2015b; Cripps et al. 2014). Is the ARIADNE project succumbing to the same myopia, which their D2.1 survey exposes? Is it failing to acknowledge the fact that Semantic Web technologies show little signs of gaining traction within grassroots archaeology? While at the same time, it promotes the views of the data managers who are seemingly happy to commit to public ontologies and other methods firmly embedded within the praxis of the Archaeological Semantic Web. In the next section, we drill down further into the social nature of the Archaeological Semantic Web by reviewing the findings of a series of workshops conducted on the subject.

Semantic Web workshops in review

In the spring/summer 2015, I conducted a series of workshops on the subject of the Semantic Web in order to better understand this techno-social dynamic. These events attracted participants from a wide range of academic and professional backgrounds, although the majority were members of the digital humanities community. The primary objective of these workshops was to assess whether it was possible to define a model of the typical Archaeological Semantic Web user. If it was the case that ARIADNE and
other large institutional bodies associated with the management of archaeological and other cultural heritage data were moving inexorably towards Semantic Web solutions to the problems associated with heterogeneous data, then I wanted to find out about the technical skills that a user would need to fully exploit these shared resources.

The first workshop was part of the ‘Data Visualization for the Arts and Humanities’ event, held over two days from 5-6 March 2015 in Queen’s University Belfast and sponsored by the Digital Arts and Humanities PhD programme.\textsuperscript{240} The workshop was attended by about ten researchers and lasted eight hours with the first half being devoted to the subject of data mining of cultural heritage resources on the Semantic Web. In the second half, the participants were instructed on how to visualise the data gathered during the first half using a Leaflet-powered web-map.\textsuperscript{241} While the data-mining component was introduced with a lecture-style overview of the key philosophical (Open Access, Open Data, Linked Data) and practical themes (RDF, SPARQL) underpinning the Semantic Web, the vast majority of the teaching took the form of practical exercises, which became incrementally more complex as they progressed. First, the participants were asked to create some RDF content, in order to familiarise themselves with the fundamental Knowledge Representation structure of the Semantic Web. The remainder of the exercises were completed using SPARQL and a sample of the types of questions asked of the participants is reproduced below.

Q: Get a list of all of the universities known to DBpedia
Q: Get a list of the Getty URIs that represents concepts related to amphorae
Q: Get the find spots of all of the sarcophagi in the British Museum collection
Q: Get the geo-coordinates of all of the coin hoards stored in the Nomisma collection

The final exercise carried out by the group was to generate a list of ships on DBpedia. Each resource needed to include a link to an image reference, label and a place name. This data was then cleaned and the place name transformed into a set of geo-coordinates using Open Refine. It was then passed on to serve as the source data for the second half of the workshop. While no quantitative analysis of the participants’ performances was

\textsuperscript{240} Details for the event can be found at http://www.eventbrite.ie/e/data-visualization-for-the-arts-and-humanities-tickets-14880944305.

\textsuperscript{241} I took charge of the teaching for the data-mining component of the event. My colleague, Vinayak Das Gupta, led the visualisation module.
carried out during the four hours of the workshop, it was clear that some participants grasped the concepts involved and techniques better than others, and the knowledge backgrounds of the participants appear to have been significant in this respect. All of the members of the group achieved the baseline requirement of completing the DBpedia ship exercise but it was evident that not all would have been subsequently able to use the concepts and techniques described to formulate new questions and to generate new knowledge autonomously (Mayer 2002 p. 226). The participants who did seem to successfully internalise the ideas contained within the exercises all shared a common foundation of scientific training.

The second workshop, also sponsored by the Digital Arts and Humanities PhD programme, replicated the structure of the Queen’s workshop in that it was again divided into two conceptual halves, albeit within a slightly more contracted timeframe. While the first audience group included participants from both science and humanities backgrounds, the majority of the ten participants of the second workshop were enrolled in digital arts or humanities postgraduate programmes. Again it was the case that no objective measurement strategy was employed to assess the participants’ experience of the workshop. In general, the participants of the second workshop found the material to be more challenging than was the case for the first group. A lower proportion of the participants appeared to make the leap towards understanding the key concepts and my suspicion is that the higher representation of humanities and arts researchers in the second workshop may have been significant in producing this outcome. It is also possible that the slightly more constrained timeframe and the lack of an intervening overnight mental break may also have played a role.

This account of the Semantic Web workshops carried out during this project should not be read as an exercise in controlled empirical study. As has been mentioned, the interpretations arrived at have been mainly based on a set of non-scientific observations. Another point that needs to be made clear is that these studies placed a number of constraints on the participants. For instance, a participant was not allowed to propose an answer using a method of their choice. They were required to work through each question by writing SPARQL queries, which were to be addressed at specific Semantic

\[242\] Details for the event can be found at [http://www.eventbrite.ie/e/16695719346](http://www.eventbrite.ie/e/16695719346).
Web resources that had been pre-identified for them. As such, the workshops imposed a very particular work practice on each of the participants.

Bearing these caveats in mind, these findings do, I believe, provide valuable information about the nature of human engagement with Semantic Web research methodologies. The workshops suggest that the SPARQL/Semantic Web research combination is a more natural fit for researchers who bring with them some background knowledge of the scientific method. This does not mean to suggest that non-scientifically minded archaeologists cannot become skilled at using the SPARQL method but rather that this group will need to invest more time and effort in upskilling than those who have this prior scientific training. Having said that, the workshops also showed that regardless of the intellectual background of the research candidate, it is necessary to invest a certain baseline time commitment to the learning of the basics of RDF and SPARQL before profitable Semantic Web research can be attempted. SPARQL’s intolerance of syntax errors and requirement that the user know the target ontology well both demand that the user invest significant amounts of time in mastering its formal constructs and the arts of data structure discovery, which at first can appear as formidable challenges.

But for most potential Archaeological Semantic Web use-case scenarios, this training requirement model would not be unworkable or indeed unusual when viewed in the wider context of digital humanities pedagogy. To prescribe that all users of the Archaeological Semantic Web first become acquainted with the basic principles involved and in the use of SPARQL over the course of perhaps ten hours of training would not be, I believe, unreasonable. After all, the TEI method, which has been rightly hailed as a success story in the application of digital methods to text-based humanities questions, is predicated on the idea that the user first spend a not insignificant amount of time in a formal learning process before they can start to produce and consume TEI materials productively.

There is, however, another reading of the outcomes of the Semantic Web workshops and that is that it might not be possible to envision the Semantic Web as a resource usable by the ‘average’ archaeological researcher. In this scenario, the Archaeological Semantic

---

243 Perhaps because the motivation is simply not yet there for the average researcher to invest in learning the necessary skillset.
Web becomes a knowledge domain, which can only be negotiated by specialised Semantic Web technicians. While this model goes against the general trend and aspiration of the digital humanities, which is in many respects a manifesto reaction against the fetishisation of technology and a tendency of humanities projects to outsource their technical needs to a separate computer science ‘tool’, it cannot be dismissed out of hand (Kirschenbaum 2009; Ramsey 2013). In that light, I conducted a final and more specialised workshop session, which aimed to further develop this idea of the Semantic Web specialised user.244

This workshop was conducted with a sample group of just three individuals, with only one of these having a background in the humanities (Classics). The participants were chosen instead because they had knowledge of existing data management paradigms (specifically SQL-based systems) and for their knowledge of computer programming. Knowledge of SPARQL was not assumed and indeed only one of the participants had used it to any extent previously. The objective of this final workshop was to investigate whether it was possible to train a technician in the use of the Semantic Web within a relatively short space of time (three hours) and whether their technical background knowledge would enable them to make the conceptual leap from learning to understanding faster than was the case with the participants of the first two workshops. As a secondary consideration, the workshop was also configured so as to compare the traditional form of query interface found on websites against the asking of queries using SPARQL. The linkedarc.net web app was used as the test environment and the data of the Priniatikos Pyrgos archaeological project as the test dataset.

<table>
<thead>
<tr>
<th><strong>Objective</strong></th>
<th>The site of Priniatikos Pyrgos was excavated and recorded from 2007 on using the single context system. In this system, all data is tied to the central context data concept. Frank Lynam excavated at Priniatikos Pyrgos from 2007-2010. I want you to find out and represent how much (weight and sherd count) pottery he excavated during this time and then to compare this against that which was excavated by David Govantes-Edwards.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Guidelines</strong></td>
<td>I want you to use the linkedarc.net web app to access this</td>
</tr>
</tbody>
</table>

244 The exercises included in the Semantic Web specialised user workshop are reproduced in Appendix C.
information. You can use any of its features except for the SPARQL interface. Once you get this information, I want you to do the same for David Govantes-Edwards. Finally, create a chart that compares the data for the two excavators.

Table 42: Semantic Web workshop 3 – exercise 1 instructions

The instructions for the first exercise are reproduced in Table 42. In general, the level of successful completion of this task was very low across the board, and in general the participants found it difficult to navigate the user interface of the linkedarc.net web app. While, it is possible (perhaps even probable) that the design of the linkedarc.net web app’s search and navigation interface was the cause of this return, the participants’ lack of any prior knowledge of archaeological concepts (particularly of the Single Context terminology employed by the Priniatikos Pyrgos test dataset) might also have been a factor.

Objective

I want you to create a timeline of the Priniatikos Pyrgos excavation by first asking questions of the linkedarc.net SPARQL endpoint and then visualising this data using the Timeline JS service. Ideally, I am looking for something like the timeline shown in the attached image.

Guidelines

I am not going to give you any initial overview of the structure of the Priniatikos Pyrgos material, which is hosted on linkedarc.net, except to say that it conforms to the English Heritage extension (CRM-EH) to the CIDOC CRM (CRM). You will need to use the linkedarc.net web app and the online guides to the CRM and the CRM-EH to work this out for yourself. At a minimum you will need to retrieve
each of the context names, their excavation date and their descriptions. If you have time in the end, you can supplement this by adding an image for each of the contexts.

Timeline JS is a really nice data visualisation platform and very simple to use. It allows you to create timeline visualisations for time-based data like the excavation dates for the Priniatikos Pyrgos contexts. The site includes a very easy to follow guide about how to get it up and running.

<table>
<thead>
<tr>
<th>Table 43: Semantic Web workshop 3 – exercise 2 instructions</th>
</tr>
</thead>
</table>

The second exercise’s guidelines are reproduced in Table 43. In this exercise, the participants were asked to use the linkedarc.net SPARQL interface to construct a query that would be able to retrieve the necessary data from the linkedarc.net triplestore. While ostensibly a more complex challenge, the level of progress achieved in this exercise was far greater than that observed during the first exercise. This does not mean to say that the second challenge was deemed straightforward by the participants. In fact, only one of the three participants managed to accomplish all that was required, while the remaining two participants rated the exercise as being either hard or very hard. But despite this perceived difficulty, all three participants were far more comfortable using the prescribed workflow than was the case for the first exercise. The construction of SPARQL queries, while initially challenging, soon resonated with previous experiences of writing analogous logical queries (mainly SQL), and the participants were able to transform this prior knowledge into the writing of quite complex SPARQL queries by the end of the 3-hour session.

**The Archaeological Semantic Web user model**

In the introductory chapter of this thesis, an attempt was made to position this work within an existing intellectual context. It was stated that this is a product of the digital humanities. Often the act of assigning labels to a work ends up being an exercise in intellectual efficacy. Labels allow the creator and the reader to easily categorise a piece of work, which necessarily determines aspects of its production and consumption. Existing knowledge caches can also provide models, which can be re-used by the author for the purposes of increasing the coherency and intellectual rigor of the final work. In Chapter 1, we defined five types of digital humanist: DHer 1, 2, 3, 4 and 5. The objective of this
The following model of the Archaeological Semantic Web user is directed specifically at this second category of digital archaeologist – hence the omission of user types analogous to the DH project manager (DHer 4) and the domain knowledge specialist (DHer 5). It models the demands that are placed on, the general working practice and

245 Take for example the work of the Universities of Southampton and Cambridge at the ancient Roman port of Portus near Ostia (Keay 2011). While the project team uses a number of digital interventions (particularly those relating to visualization), the core work practice remains essentially traditional (field survey and excavation).
the output types that one might expect of the digital archaeologist who works in this particular social environment. In this scenario, our target archaeologist is first and foremost trained in the study of past material culture. However, like many archaeologists (physical anthropologists, ceramic specialists, geo-archaeologists and so on) they are also skilled in highly specific sub-fields of the discipline, which in their case is the application of Semantic Web technologies to the problems of structuring, analysing and representing of archaeological datasets.

These are the parameters that need to be borne in mind when reading the following model. The ASWer model should not be misconstrued as an attempt to define all forms of interaction with the Archaeological Semantic Web for it clearly could not defend such a claim. As with all models and following McCarthy’s lead in the DH context (McCarty 2004), it should be seen rather as an interpretive aid with which to better understand the form and habitus of Archaeological Semantic Web engagement.

**ASWer 1 – baseline Archaeological Semantic Web user**

The first type of Archaeological Semantic Web user, which we abbreviate as ASWer1, is analogous to DHer 1 in the sense that they are the baseline user of the Archaeological Semantic Web (Figure 164). ASWer 1 has sufficient technical knowledge to be able to create SPARQL queries of simple and intermediate levels of complexity, which enables them to ask questions of most resources on the Archaeological Semantic Web. They are also familiar with the apparatus of the RDF triplestore infrastructure from a consumer perspective. This means that they know how to browse through RDF data and to request this data in various serialisations: JSON-LD, RDF Turtle and RDF/XML. They are also

---

**Figure 164: the three Archaeological Semantic Web user types**

ASWer 1
- baseline user
- 10 hrs of training in SPARQL, RDF, ontology discovery
- 5 hrs of training in data cleaning, geo-coding

ASWer 2
- advanced tool user
- tools: Gephi, GIS, R, ...
- no coding or basic coding

ASWer 3
- tool creator
- know how to code
- domain knowledge

+ ASWer 1
+ ASWer 2
able to interpret these serialisations. ASWer 1 will have the skills to discover the datasets on the Archaeological Semantic Web that might be relevant to their research questions. ASWer 1 will also know how to negotiate the structure of a previously unknown dataset.

ASWer 1 would need to complete ten hours of instruction in basic Semantic Web practice in order to acquire the skillset outlined above. Ideally, this would be supplemented with an additional five hours of instruction in the use of a data cleaning and post-processing application, such as Excel or Open Refine. This would allow the ASWer 1 to clean up retrieved data, thereby increasing its subsequent research value. This second phase would also teach them how to augment the data using techniques such as geocoding using the Google Maps Web API. Having completed the combined 15 hours of training, the ASWer 1 will be able to comfortably construct and pose questions of relevant Archaeological Semantic Web resources and use the results of these queries to supplement other core research methodologies.

**ASWer 2 – advanced Archaeological Semantic Web tool user**

The ASWer 2 user has all of the skills of ASWer 1. ASWer 2 is like DHer 2 in that she combines advanced tool knowledge with knowledge of SPARQL, RDF and various commonly used data models and vocabularies. The particular tool used by the ASWer 2 user will depend on their particular research interest. If, for instance, their research were primarily focussed on subjects of a geospatial character, it would make sense for them to apply their Archaeological Semantic Web knowledge to a GIS tool such as ArcGIS or Quantum GIS. If networks were their interest, then a tool like Gephi could be used, as is described in Chapter 6. A statistical focus could be accommodated using the R software platform (Carlson 2012). The difference between ASWer 1 and ASWer 2 is manifested in the different characters of their outcomes. Whereas, ASWer 1 predominantly deals with the data derived from their Archaeological Semantic Web data mining as is, ASWer 2 subjects this data to a secondary level of interpretation, be that in the form of statistical analysis, geospatial analysis or some other form of secondary visualisation. Kitchin would call this new data, ‘tertiary’ data (Kitchin 2014 p. 8). The ASWer 2’s primary expertise is in the use of the core tool. Their Archaeological Semantic Web skills constitute supplementary knowledge, which should be employed as and when the situation dictates that they will be of interpretive benefit.
ASWer 3 – Archaeological Semantic Web maker

The final Archaeological Semantic Web user type proposed by this model is known as ASWer 3. ASWer 3 has a high degree of knowledge and experience of technology in general. They know how to write code and to use this code to build relatively sophisticated archaeological solutions that consume data derived from the Archaeological Semantic Web. Following the DHer 3 model, they are tool builders, in contrast to ASWer 1 and ASWer 2, who use tools of varying degrees of complexity (Ramsey 2013). The PPTextiles case study outlined in Chapter 6 is an example of the sort of output produced by ASWer 3. What differentiates ASWer 3 from a computer scientist is their domain knowledge. They must have a research background in archaeology and, as such, they need to be entirely comfortable with conceptualising archaeological information. For instance, in the third workshop account included above, it was noted that the three participants were at a disadvantage because they did not have any background experience of dealing with archaeological concepts. The creation of tools by developers who have that sort of knowledge is not the norm. It is far more common to find archaeologists using tools that have been created by generic problem solvers working with generic problems in mind. Generic approaches must necessarily be inferior to approaches that have been designed with particular problems in mind.

If there is an aspiration within archaeology to start designing Archaeological Information Systems that truly embrace the potential offered by the digital medium, then it will be necessary to ensure that archaeologists are involved at all stages of the development, including the writing of code.\textsuperscript{246} This would necessitate a fairly significant paradigm shift in the way that digital tools for the humanities have been developed to date following a model, which has relied heavily on the outsourcing of the technical roles to computer scientists. If digital archaeology and digital humanities in a more general sense are to achieve higher levels of digital engagement by moving beyond the mimicry of pre-digital practices, then it will be necessary to forego the outsourcing model by increasing the numbers of DHer 3s and ASWer 3s active within the field.

\textsuperscript{246} This is similar to but not exactly the same as relying solely on user-centred design in conjunction with a strong collaborative project philosophy. While the latter can go some way towards bridging the tool creation/domain knowledge gap, it is my firm belief that this strategy will never be able to fully eclipse that gap in the way that their combination can.
The epistemological implications of the ASW user model

If we accept this model for the use of the Archaeological Semantic Web, then we must also consider its implications, particularly those that relate to the matter of archaeological epistemology. In order to do this we first need to have a firm grasp of our terminology; and so we begin this section with an examination of epistemology from a classical and then a digital perspective. We follow this by considering how well this digital epistemological model accommodates our tripartite Archaeological Semantic Web user framework. And lastly, we consider how our model fits within the framework of existing archaeological schools of thought.

What is classical epistemology?

In order to address the question of digital epistemology, we need to first consider what it is that we mean by epistemology in a traditional or 'classical' sense (Dede 2008). A good place to begin would be to consider how epistemology might relate to ontology, which is a concept with which we should now be fairly familiar following the discussion contained in Chapter 3. If the ontology is a model of reality or a domain of reality, then epistemology is the study of how we access that model. If we manage to access the ontology correctly, we are said to obtain episteme or truth. As such, epistemology is very much related to the accessing of answers through a process of questioning. When we ask who, what, when, how, and why, we are engaging in an epistemological act, in that we are attempting to discover an acknowledged truth of a matter.

![Figure 165: knowledge pyramid (Kitchin 2014 fig. 1.1)](image-url)

---

247 *Episteme* is an ancient Greek word used by Plato to convey a meaning that we would now define as truth (Parry 2014). It is contrasted with *doxa* or opinion.
Philosophers have been attempting to break epistemology down into its component parts since the time of Socrates, so as to better understand how it is that we know what we know (Hintikka 2007). One frequently used model with which to consider the construction of knowledge is to imagine that it is structured in the form of a pyramid (Figure 165). This vertical and hierarchical perspective is very much rooted in a conception of Knowledge Acquisition as a highly structured process; it is one which gained great traction during the Age of Enlightenment and subsequently as a codified didactic model in the German University system during the 19th centuries (Berry 2012 p. 7; Liu 2014). At the bottom of this pyramid is the phenomenal world and as you move progressively upward, you pass from layers of higher to lesser degrees of abstraction (Kitchin 2014 p. 9). This passage is made possible by the application of different types of human cognitive function. By the time you reach the pinnacle of the pyramid, you will have arrived at the application of knowledge, which we know as wisdom.

The emphasis here is on obtaining progressively higher forms of knowledge through a process of cognitive distillation. This is essentially an Aristotelian view of knowledge, which is highly anthropocentric (Berlin 1982 p. 767). As a combined process, it is crystallised most purely in the form of the traditional scholar’s practice. This individual has a number of characteristics. They are likely to work alone, and for them a problem’s truth is represented by a single explanation (Dede 2008). As such, duplicate simultaneous explanations are not possible and must be corrected as errors (Lynch 2011 p. 3). The medium of knowledge and the communication of that knowledge are also very important. We have already discussed in Chapter 2 how technological changes have had vast social, political and economic effects. Before the invention of writing, the oral method was the only way that humans could exchange internalised knowledge (Hyman & Renn 2012 p. 8). Cline argues that the arrival of writing and its emancipation by Gutenberg’s invention of the printing press changed knowledge from an oral process to a primarily visual endeavour – a statement of truth transformed from ‘people say’ to ‘it has been observed’ (B. Cline 2012 p. 7). Truth, therefore, in all epistemological eras, is as much to do with authority and trust as it is with obtaining the ‘real’ fact of the matter. In this context, professional knowledge practitioners came to be trusted, not just because they could write and publish but also because of the way that they did this. Value was placed on publications that used high language registers and those that diverged from this practice came to be frowned upon (Liu 2014).
Ultimately, classical epistemology is characterised by its linearity of approach, fixity, conservatism, and resistance to change (Rorabaugh 2012). Institutions are also a core element of the classical epistemological system. The dominant social institutions of religion and the state have always maintained a vested interest in preserving the societal status quo, and the management of knowledge in centralised knowledge institutions has proved instrumental in ensuring that none of the ‘wrong’ sorts of ideas were promulgated (Carl 2009 p. 503). Even with the arrival of the Age of Enlightenment and with all the seismic changes that this brought about, the fundamental structures of epistemology remained largely the same. Knowledge was something to be produced and consumed in certain centres of learning, and in more recent times a carefully controlled version of it was to be disseminated to the populous in the form of state schooling.

**What is digital epistemology?**

Jean Francois Lyotard proposed that the epistemological changes described above came about as a result of parallel social, cultural and economic shifts (Lyotard 1984 pp. 3–4). Lyotard also suggested that, while these effects have always influenced the nature of our conception of and approach to knowledge, the advent of modernism resulted in a more profound transformation. With capitalism as the driving philosophical force and mass media as its technological conduit, knowledge became a commodity to be traded and exchanged in the modern world. In other words, ‘knowledge ceases to be end in itself, it loses its “use-value”’ (Lyotard 1984 p. 5).

This post-industrial revolution intellectual climate reigned for about two centuries and, in that time, the scale of the commoditisation of knowledge expanded, but for the most part the nature of the knowledge system remained largely static. All of this came into question, however, with the dawning of the digital era, which in the eyes of many commentators instigated epistemological transformations of a type never before witnessed (de Bolla 2014; Dede 2008; Liu 2014; White 2014). To begin with, the medium of knowledge had changed and we know from our review of classical epistemology that media changes can be very significant for knowledge – in fact, if we follow McLuhan (McLuhan & Fiore 1967), the medium is not so much subordinate to the message, as it is the message.248 The knowledge producer in the digital era is no longer constrained by the

---

248 Baudrillard’s (1994) hyperreality model is an extension of this idea.
physical form of the printed page. Their canvas is a far broader and flexible space than was the case in the pre-digital era.

One of the more fundamental and ubiquitous digital innovations is what is known as hypertextuality, meaning the ability to link elements within different document resources (Cicconi 1998; Oblak 2005). Cline believes that the applications of hypertextuality – footnote linking, linking to relevant external resources and definitions – render the e-reader, as an example, an entirely different epistemological animal to the print book (B. Cline 2012 p. 9). For Cline, e-reader epistemology is like all forms of epistemology – it is not a thing but a process – and the process of consuming digital media using the e-reader is substantively different to that of reading a book.

Another notable aspect of digital epistemology cited by various scholars is its apparent state of flux (Jenson & de Castell 2012 p. 4), which is no small part due to the fundamentally different knowledge production/consumption environment that it has fostered. As was noted in Chapter 2, Web 2.0, or the Social Web, cast aside centuries of established knowledge production and consumption norms in the space of less than a decade. Truth is obviously a central idea of any epistemology. In its correspondence theory understanding, it is a proposition that reflects an underlying reality (Glanzberg 2014), which is accepted by a majority of the domain’s population (Babbie 2010 p. 37). In classical epistemology, truth is obtained through a slow iterative and linear process. Science codified this process but it did not necessarily speed it up. It takes time to observe a phenomenon and induce a theory from that series of observations. Web 2.0 turned this model on its head by allowing truths to come into being almost instantaneously with the typing of a tweet or the posting of a Facebook status update. And in the same breath, they can evaporate as quickly again, hidden beneath a constant deluge of new digital content.

What, on first inspection, are apparently more lasting forms of content, such as the articles hosted by Wikipedia, turn out to be no less transitory in reality (Evans & Rees 2012 p. 25). Entries in this vast digital knowledge store change by the day as and when the service’s army of editors decide that a change is to be made. They are, as a result, perpetually in a state of ‘under construction’. This introduces another important aspect of digital epistemology and that is collaboration (Edmond 2016). Digital knowledge is,
more often than not, produced as an output of a collaborative process. Digital content creators form informal, decentralised networks, which construct digital knowledge for no reason other than an apparent desire to contribute to the community (Klisanin 2011). Another peculiar aspect of the Wikipedia model is that it has never been monetised and because of the network’s popularity, this has effectively meant that general knowledge of the sort that Wikipedia curates will remain a non-commercial suburb of digital culture for the foreseeable future (Mason 2015).

An outcome of this apparent increased ‘transience’ of truth is a heightened scepticism on the part of the consumer (B. Cline 2012 p. 10). As knowledge tends not to be subjected to the forms of interrogation and study that it once was, it is natural that consumers will be more wary of the veracity of the material that they encounter in the digital arena (and indeed outside of this realm as well). Essentially, this is matter of trust and it is one that has been associated with epistemology since time immemorial. In fact, I would goes so far as to contend that epistemology is as much a study of an information transfer’s authority as it is about the pursuit of ontologies or truth (Origgi 2008). Digital knowledge services such as Wikipedia have changed the parameters within which information is allowed to flow but trust has remained as critical to their operation as it ever was in the print era (Adler et al. 2008; Javanmardi et al. 2009; Lucassen & Schraagen 2010). One need only look at the design of Wikipedia and the effort that it makes to render transparent the authorship history of each of its digital resources to realise that it takes the authority challenge very seriously. If a knowledge consumer does not trust a digital resource, than they will not use it. How to gain the required level of trust is complex and largely built upon structures (footnotes, peer review, academic affiliations, the structures and methodologies applied, metadata, paradata) that have been established within the print knowledge world over centuries (Warwick 2012).

How a digital resource comes to have authority presents a very particular challenge when it comes to dealing with Open Data digital resources. These resources are by their nature more exposed to the possibility of compromise than their closed counterparts. Trust, therefore in the Open Data context, is a factor for both sides of the knowledge equation. First, as we have shown, the knowledge consumer needs to be reassured that the digital resource that they wish to exploit is imbued with a satisfactory level of authority. However, the same concern is at play for the knowledge producers (individuals or, more
probably, institutions in the Semantic Web context). From their perspective, they need to know that their intellectual property will retain its authority once it has been released as Open Data and altered as a function of that legal and technological character shift (Zorich 2008 p. 28).

Linked to the concepts of the transience of truth and the principle that digital information may be less authoritative than its print equivalent is the idea that any one reality can have more than one simultaneous truth (Hyman & Renn 2012 p. 14). Even before the arrival of digital representations, postmodernists were espousing the value of such an epistemological position, and, ultimately, this probabilistic view of knowledge can be traced to the invention of quantum mechanics. In quantum mechanics, any single phenomenon can have more than one state at a point in time, as a function of it being observed (Heisenberg 1958 pp. 46–47; Josephson & Josephson 1996 p. 1).

In recent times, a number of scholars have chosen to view digital epistemology through a materialistic lens (Liu 2014; McCarty 2008; Ramsey & Rockwell 2012). Digital publications in this model are seen as material objects, which is an attractive metaphor for a number of reasons. First, it suits the related view of digital humanities as a field essentially concerned with the act of making, which we introduced in Chapter 1. In this sense, digital epistemology becomes a type of craft knowledge, which is knowledge that has been produced using non-traditional and consciously practical means (Kristina & Townsend 2010). This implies that coding is analogous to writing in the sense that it represents the logical working through of a problem, not unlike the function that a prototype or experiment in science performs (Ramsey & Rockwell 2012). Hyman and Renn’s (2012) appropriate the creator paradigm when proposing their ‘prosumer’ model for digital knowledge consumption. Their prosumer behaves simultaneously as a producer and a creator of knowledge, and this is facilitated by the use of new knowledge exchange platforms that promote collaborative work and bidirectional knowledge trajectories (Hyman & Renn 2012 p. 14).

The remaining facet of digital epistemology, which we have yet to address, is the matter of scale, both in terms of the amount of information floating about in the digital epistemological space and also in terms of the quantum of people that are involved in these transactions (Hayles 2012 pp. 45–46). To an extent, all or most of the facets of
digital epistemology have come about, at least in part, because of this phenomenon. Liu proposes a model for digital epistemology, which explicitly references digital scale in both of the senses that we have described it. He imagines this new epistemology as being flat and horizontal, in contrast to the tower-like hierarchy of classical epistemology, which is manifest in the form of the knowledge pyramid (Liu 2014). This epistemology is flat because its production base is vastly larger than has been the case at any previous point in the history of knowledge. He calls this ‘the wisdom of the crowd’ and the journey from producer to consumer no longer involves the same level of processing that is a characteristic of classical knowledge. It is ‘inch-deep but mile-wide’ (Figure 166).

Accommodating the ASWer model within digital epistemology

How then might our Archaeological Semantic Web user model fit within the epistemology described in the last section? Remember that we defined the Archaeological Semantic Web user along three lines. The ASWer 1 user is the baseline candidate. Her principle technical competence is an ability to write SPARQL queries and to interpret the responses to those queries. ASWer 2 can do all that ASWer 1 is capable of and this is supplemented with an expert knowledge of an advanced tool in which to feed the data mined from the Archaeological Semantic Web. Finally, ASWer 3 creates tools, which consume Archaeological Semantic Web services.

It is tempting to consider all three of these user types as being consistent with the agent inferred by Liu’s model. They create and also consume data on the Archaeological Semantic Web and their publication of data as knowledge reaches other consumers extremely quickly. ASWer 3 is an ideal fit for Ramsey and Rockwell’s model of the digital humanist as creator and Hyland and Renn’s prosumer model also appears promising for all three ASWer users. However, it is here that a problem begins to emerge. Given that the Ramsey/Rockwell and Hyland/Renn models and all three categories of Archaeological Semantic Web user assume as a prerequisite a relatively high level of
specialist knowledge, does this not go against the basic principles of Liu’s more radical democracy of opinion model for digital epistemology? In fact, it is not much of a leap to instead view the Ramsey/Rockwell and Hyland/Renn models as tending more towards the exclusivity and rarefication more consistent with classical epistemology than an epistemology based on the radical knowledge freedoms of Web 2.0.

If we return to the findings of the ARIADNE user needs survey discussed earlier in this chapter and we recall the dichotomy that emerged between the responses of archaeological researchers with those of the data managers on the subject of control (ontologies and controlled vocabulary use), a pattern starts to emerge. My hypothesis goes as follows. To use the Archaeological Semantic Web demands a reasonably high level of expertise. As prescribed earlier in this chapter, this knowledge is best acquired in a formal learning environment following the model of TEI training that is already a well-established facet of most digital humanities didactic environments. The second point is that any humanist can be a user of the Archaeological Semantic Web, if they go through this training, but as our workshop data appears to indicate, candidates with an experience of the scientific method are more likely to excel with this challenge. Thirdly, once trained, users of the Archaeological Semantic Web will most likely be required to use sophisticated digital techniques as part of their Semantic Web interactions, which may or may not include coding. Fourth and finally, all of these requirements necessarily mean that Archaeological Semantic Web use is the preserve of the few and because of this, it is a method more consistent with classical epistemological environments than the digital epistemology described by the likes of Liu. And by the same logic, the Archaeological Semantic Web as a knowledge production medium is similarly compatible with classical epistemology. The accounts contained in Chapters 4 and 5 have clearly illustrated how creating RDF content for publication to the Archaeological Semantic Web places enormous demands on the knowledge creator.

**On the possibilities and challenges of the Archaeological Semantic Web**

It still remains for us to comment on the nature of the more specific form of epistemology that the Archaeological Semantic Web allows for and indeed promotes. In this next section we consider a number of the characteristics of the model and ruminate on their possible epistemological implications.
Data linking and dynamic knowledge systems

Whenever I have been asked to talk on the subject of the Archaeological Semantic Web, I have tended to begin with a vision of an archaeological future, which has fully embraced the ideals and practices of the model. In this imaginary world, data and interpretation are published in parallel, rendering both entirely transparent. This new knowledge environment is, therefore, completely open – any data source can be accessed and interrogated in it. As well as this, all of the data that constitute the nodes of this future Archaeological Semantic Web’s graph are linked to the data of other datasets. In that sense, it is better to speak of a Semantic Web, because in this future archaeological datasets will link to data, which we now associate with other disciplines. This knowledge will not be inter-disciplinary or multi-disciplinary. These epistemological categories, which have become a staple of pedagogic discussions in today’s world, will cease to satisfy any practical or theoretical need and will in a sense become meaningless.

I end this dream of a future Semantic Web with the coup de grâce. In this fully linked knowledge domain, all knowledge will be entirely dynamic. To explain this concept, I outline the following scenario. Imagine the publication of the findings of two archaeological investigations, which are in some way linked, perhaps through the typologies used by archaeologists to classify their material assemblages or through shared historical trajectories. We will call the publisher of the first investigation, Agent A and the publisher of the second, Agent B. Agent A publishes a dataset and an interpretation of that dataset in year 1 of our scenario. Because the interpretation is linked to the data and because the data is linked to other data, as and when these supporting datasets change, as will inevitably happen, these effects will be reflected in the interpretation. By year 3, Agent A’s account of their excavation on the island of Crete presents different results to its year 1 representation. This is because certain assumptions contained within the account are based on links to the datasets of other Cretan datasets. For example, Agent A discovered an imported ceramic vessel while excavating at their site. This import comes from the site excavated by Agent B. Agent B believed that this particular pot type was dated to the Early Archaic period and this interpretation had implications for Agent A’s interpretation. By year 3 Agent B had changed their mind about the date of the pot type. They now believe it to be Late Archaic or even Classical in date. In the dynamic epistemology of linked knowledge that is described in this fantasy, Agent A’s interpretation changes automatically, without the need for any human interference.
Clearly, this is a rather radical reimagining of the way that a truly global and fully linked knowledge system could work. And yet, when considered, it is exactly analogous to the way that the dissemination of new ideas happens today and has happened for thousands of years, except that the timeframes involved have been collapsed and the scale and depth of inferences expanded. A common feature of the work practice of all archaeologists, and indeed all humanists, is that they build upon the work of others. The citation system that dominates the traditional form of scholarly dissemination is predicated on this notion of linking to existing knowledge (Hellqvist 2009). We have already discussed how scholars of new media have heralded hypertextuality as the logical mechanism with which to port this system over for use by document web resources, but to date there has been little comment about the epistemological implications of applying the same thinking to data.

In theory, the deceptively simple form of the RDF triple in conjunction with the ubiquitous HTTP protocol stack make a lot of this futuristic-sounding potentiality technically possible today. There are, however, two problems. We noted in Chapter 5 how Semantic Web resources have in general failed to garner the level of trust needed to encourage linking between datasets. And reports about the macro state of the sector by groups such as Linked Open Data Project suggest that we are still a long way off achieving the degree of data linking that consumers of the document web take for granted (Schmachtenberg et al. 2014). The second problem is related to the matter of indexing across multiple datasets and we will discuss this next.

**On finding data on the Archaeological Semantic Web and data myopia**

Currently, there is no Google-like integrated technological system that allows the entire Semantic Web to be searched as a single entity. What this means is that if a data consumer wants to find out about a certain artefact type from a particular geo-temporal context, for example, they will need to specifically target each Archaeological Semantic Web resource that might contain data related to this set of criteria. This means that the researcher will need to first compile a list of these resources and then interrogate each in turn using what may be entirely different search mechanisms (Paulheim & Hertling 2013). This presents the distinct possibility that the researcher will either omit a relevant Archaeological Semantic Web resource from their initial list and/or fail to find each relevant data node in the resources, which are interrogated, as is illustrated in Figure 167.
This has significant epistemological implications, as when working in such an environment, the consumer cannot always be sure that they are accessing the full breadth of data resources relevant to their question. Of course, this idea of the impartial source view has been an inherent characteristic of all epistemological endeavours to date. The responsibility of the traditional scholar has been to ensure as far is possible that their partial data myopia is attenuated by constantly staying abreast of the latest research in their field and by being skilled in the data-finding method. This latter requirement has traditionally meant that scholars have been required to know how to use library and archive facilities efficiently and intelligently (DeMarrais & Lapan 2003), and while the context of the enquiry might have changed for research carried out on Archaeological Semantic Web resources, the principles remain largely the same. In Chapter 8, we propose a possible solution to this problem for the Archaeological Semantic Web.

**Epistemological determinism**

Determinism and free will are inextricably linked concepts. Free will is defined by the *OED* as ‘spontaneous or unconstrained will; unforced choice; (also) inclination to act without suggestion from others’ and determinism in a general sense as ‘the doctrine that everything that happens is determined by a necessary chain of causation’. In broad terms, there are three possible philosophical positions that can be adopted towards these
concepts (O’Connor 2014). The first position states that if free will exists in any sense, this negates the possibility of determinism’s existence. Second, there is the model that both determinism and free will are terminals on a spectrum. This implies that the more that you possess of one, the less you have of the other. Lastly, it can be argued that all human action is determined and that free will is an aberration.

While technologists (and indeed archaeologists (Huggett 2012 p. 205)) have tended to avoid considerations of the deterministic effects of their systems, the causal nature of technology has not been universally ignored (Dusek 2006; Ferré 1995; Heidegger 1977 pp. 3–35; McLuhan 1977). In this context, the technological determinism argument can be reduced to whether you believe that a technological material entity can have agency and whether this can have a causal impact on the agency of humans (Ferré 1995 p. 129). Although we have argued throughout this thesis that the Archaeological Semantic Web is a flexible solution to the problem of heterogeneous archaeological data, it can also be argued that any Knowledge Representation system must necessarily employ certain practices or rules, which are likely to have deterministic outcomes.\footnote{See Faulkner and Runde’s (2009 p. 446) association of rules with behavioral causation in humans.}

Taking RDF as an example, one might assume that such an apparently simple Knowledge Representation system would be for the most part deterministically benign. However, I would argue that the RDF triple imposes a particular model of structuring knowledge that is epistemologically leading (Klobucar 2010). The triple establishes formal links between entities and chains of these links can be used to produce reasoned inferences, as we have seen in Chapter 3. It is, however, possible to imagine an alternative to this conceptual structure. For example, given the particular circumstances it might be preferable to think about entity associations built upon a hierarchical structure. In this case, in the place of a graph triple, we would need to employ a Knowledge Representation such as XML, which would enforce the rules of hierarchical relationships. The fact that a child of an entity cannot also act as its parent, which is a rule of XML, is epistemologically meaningful, because it determines the parameters of how data modelled by the Knowledge Representation can be structured. A similar argument can be made for the use of tuples in NoSQL datastores.
And then there is the matter of the public ontologies and controlled vocabularies that provide so much of the actual functionality of the Archaeological Semantic Web. The CIDOC CRM and its event-based model of reality is certainly a very particular way of viewing a knowledge domain. As was discussed in Chapter 3, the CRM might instead have been designed around an object-centric model. The fact that it chose not to is a function of a whole raft of contextual elements that surrounded its creation (the intellectual beliefs and backgrounds of its creators, technical considerations, external feature requests from vested interests, and so on) but the important point from an determinism point of view is that the use of the CRM imposes these worldviews onto the data producers and data consumers that choose or are obliged to use it.

Controlled vocabularies are arguably more open to the threat/promise of epistemological determinism. For example, when annotating a digital record representing a particular physical artefact, it might be the annotator’s wish to associate it with a particular term. If the Archaeological Information System controls these values using a vocabulary, then the annotator might have to compromise by selecting the term that is deemed to be the closest semantic match (Manovich 2012 p. 258). As was discussed in Chapter 5, the value of greater intellectual freedom is weighed up against the utility attained through the use of controlled vocabularies (they allow for better indexing, for instance).

**Accommodating different archaeological philosophies within the Archaeological Semantic Web**

Archaeology, for its part, has endured its own periods of philosophical upheaval, especially since the advent of the so-called ‘New Archaeology’ approach in the 1960s (Trigger 1996 p. 1). As has been commented upon on a number of occasions in this text, the post-processualists argued for a far more limited understanding of the past than was promised by the New Archaeologists (Hodder 1982, 1995; Shanks & Tilley 1992; Tilley 1990, 1994; Trigger 1996 p. 386). Its advocates worked throughout the 1980s to develop a consistent framework, on which archaeologists could go about the business of practicing archaeology. While post-processualism never managed to establish the uniformity of opinion and approach that the processualists achieved (and for that reason, it is better to speak of post-processual archaeologies, rather than a post-processual archaeology), certain methodological features did, nonetheless, emerge as common intellectual traits of the group. These include, but are not limited to, a welcoming of
interpretive doubt, multiple viewpoints, reflexivity, contextual analyses and a general scepticism. Post-processualists are also more likely to consider aspects of past life (gender studies, social inequality, identity, et cetera), which were often neglected by the more generalist narratives favoured by the processualists (Canuto & Yaeger 2000; Díaz-Andreu et al. 2005; Gilchrist 1997).

We have already encountered many of these methodologies and subject interests in our review of digital epistemology earlier in this chapter, and, as such, it is tempting to see post-processualism as a natural bedfellow of the Web 2.0 epistemological framework. If we accept this link, it then stands to reason that the epistemology of the Archaeological Semantic Web would be more compatible with that of processualism, given the former’s tendency towards classical epistemological patterns of practice. The question that we will briefly consider now is whether it is possible to also accommodate intellectual frameworks, which are more consistent with post-processualist thought, when using Archaeological Semantic Web digital methods.

To begin with, we would first need to address the matter of multivocality, which is the concept that sees archaeological interpretation and interpretive approaches as capable of having more than one ‘voice’ (Hodder 2008; Tilley 1994 p. 4). While conceptually this is a relatively simple idea, most current Knowledge Representation systems would struggle to deliver this kind of flexibility. For the most part, Archaeological Information Systems are designed to represent a ‘perfect’ view of any one set of archaeological circumstances (Cripps 2013 p. 487). For example, a relational database in its usual configuration allows only a single value to be stored at any one time in a record’s field. As and when the interpretation of this value changes, the current value must be erased and overwritten with the new value. When viewed from a post-processualist point of view, this is clearly problematic, as it denies the legitimacy of the previous values. There is also the possibility that a post-processualist might wish to include more than one value for a field at any one time, perhaps because they wish to represent an interpretive disagreement over the value of a record’s field. In its usual configuration, this type of functionality would not be possible using an RDBMS.
Slowly, we are beginning to see the introduction of information systems into archaeology, however, which are capable of accommodating multiple interpretations of the one referent. This ‘versioning’ of data is now a well-established branch of information system management (Sink 2011 pp. 1–2) and its principles could be applied with relative ease in an RDF-based Archaeological Information System configuration (Graube et al. 2014). A data producer could, for example, use multiple triple branches to model the same type of interpretive act more than once. Figure 168 shows an RDF graph in which the interpretive act of assigning a time period to an artefact is modelled using the event-based CIDOC CRM. As you can see, the ecrm:P39_measured predicate is used twice by the <http://example.org/artefact1> subject. Each branch of triples, which grows out of this link, represents a single period assignment act. In this case, the RDF triples model a dispute over the artefact’s correct date.

This modelling of multiple viewpoints entails a subtle but significant shift in the understanding of what the data of an Archaeological Information System actually represents. The processualist view is that this data is a digital representation of an empirical observation (McPherson 2015 p. 487). When used in a post-processualist context, however, the data is representative of an interpretation, which is contingent upon a number of different (social, epistemological and other) parameters (Cripps 2013...
p. 488). The other interesting outcome of using the event-based system employed by the CRM is that it shifts the emphasis away from object and on to process and agency, both of which have enjoyed high levels of use as theoretical devices within the post-processualist community (Dobres & Robb 2000; Dornan 2002).

Lastly, as is the now well-established practice of this thesis, we also need to comment upon this post-processualist Archaeological Semantic Web from a consumer perspective. In this context, the Archaeological Semantic Web seems much more clearly aligned with the principles of post-processualism. The central idea of the Semantic Web, that data be made available alongside the interpretation that it produces, opens up a vast plethora of new methodological possibilities to the archaeological consumer that the world of digital data silos and ‘closed’ data necessarily denies. Essentially, Open Data injects transparency into a discipline's digital practice and this should in theory lead to higher levels of critique, dialogue and reflexivity, all of which are solid post-processualist ideals. SPARQL too would seem to present opportunities for the post-processualist consumer. Despite its superficially processual form, SPARQL, as we have seen, applies substantially fewer constraints to the consumer in terms of the types of questions that can be asked. A freer consumer/data interaction should allow for more interesting perspectives to be taken on datasets, as was discussed in Chapter 5's account of the Priniatikos Pyrgos project’s ‘social’ data.

**Conclusions**

The role of this last chapter has been to pull back from the functionalist and decidedly optimistic view of the Archaeological Semantic Web cultivated in the preceding chapters, by casting a more consciously critical eye on the model in a technological, sociological and epistemological sense. The chapter began by comparing the model against a sample of its technological alternatives on the premise that all technological choices assume the rejection of alternatives. It was found that while a case can be made for the attractiveness of using the FileMaker, relational database and NoSQL systems within Archaeological Information System infrastructures, the specific matter of heterogeneous archaeological digital data is best accommodated following the Archaeological Semantic Web approach. Having said that, this chapter also acknowledged that the future of Archaeological Information System designs must, as it does today, assume that at any one time archaeologists will adopt a hybridity of approaches. In that respect, it is important to be aware of the characteristics of these alternatives, even if one is committed to pursuing a
technological and philosophical agenda more in keeping with the Archaeological Semantic Web.

The level of future success to be achieved by the Archaeological Semantic Web will be determined as much by non-technical factors, as it is by technical considerations. In this context, this chapter also presented and analysed the findings of recent research into the reception of the ideals that underpin the Archaeological Semantic Web by its user base. We first looked at the findings of two ARIADNE surveys carried out to reveal the opinions of users on a range of topics falling under the umbrella term of archaeological digital data services. The results obtained relating to the matters of digital data sharing and openness are particularly interesting, as they show clearly a disparity between the views of regular archaeological researchers and those of professional archaeological digital data managers. The picture that emerges sees the data managers as being far more accepting of the ‘costs’ that come with Archaeological Semantic Web adherence than is the case for the archaeological researcher. While the researchers want the benefits that are promised by the Archaeological Semantic Web, it appears that the majority are reluctant to be compelled to employ the use of standardised ontologies and vocabularies of terms. This general picture was further confirmed by the results of a survey that was carried out as part of the linkedarc.net project concerning attitudes to digital data sharing within the community.

It was then necessary to consider what implications these findings might have in a broader sense. In that context, I conducted a series of workshops on the subject of the Semantic Web as a resource of the digital humanist. These workshops suggested that the primary reason that archaeologists and other researchers interested in accessing cultural heritage-themed digital resources have been reticent in adopting Semantic Web research methodologies is because of a general confusion about the objectives and structure of the model and because of a perception among the community that it is difficult to ‘learn the ways’ of the Semantic Web. This position is in keeping with the conclusions of a number of previous studies on the subject (Carver 2013; Isaksen 2011). By working with the participants of the three workshops and observing their responses to the material covered and the exercises assigned, I have proposed a tripartite model of the Archaeological Semantic Web consumer. ASWer 1 is the baseline Archaeological Semantic Web user. They are skilled in the writing of SPARQL queries and in the
interrogation of datasets, whose structure is previously unknown. They are also capable of processing by cleaning, aggregating and looking for patterns within the output of these queries. ASWer 2 can do all that ASWer 1 is capable of doing, but what distinguishes them from their ASWer 1 counterpart is their advanced knowledge of a specific digital tool, such as a GIS, a visualisation or a statistics application. The idea is that they are able to bridge the gap between this tool and the data housed on the Archaeological Semantic Web. Their tools are used to analyse, visualise and generally add value to the fruits of their Archaeological Semantic Web data-mining labour. ASWer 3 again has all of the skillsets known to ASWer 1, but they use this knowledge to create digital tools for the wider community. This last Archaeological Semantic Web user is a digital humanist in the sense that Ramsey understands the type, as a creator (Ramsey 2013).

The final section of this chapter broadened the narrative focus once again in order to consider what this model might mean in the context of 21st century knowledge systems in a general sense and then archaeological schools of thought in particular. The reason for looking at the epistemology of the Archaeological Semantic Web is that the ASWer 1, 2 and 3 model is undoubtedly restrictive in a number of senses, and restriction is not a characteristic usually associated with digital epistemologies. Because of this, I conclude that the ASWer model has more in common with classical epistemological mores than with those attributed to Web 2.0. In an effort to explain this apparent paradox, I would argue that this correlation is entirely consistent with the endorsement of the Archaeological Semantic Web by the majority of data managers in contrast to the general reticence with which the researcher community approaches it. The logic here is that data managers have more of a vested interest in the maintenance of the epistemological status quo, while archaeological researchers are less concerned with these motivations. Lastly, I also argue that while, for all of these reasons the Archaeological Semantic Web sits most naturally within a theoretical framework that follows the tenets of archaeological processualism, this does not mean to say that the Archaeological Semantic Web cannot or should not be used by archaeologists who are more drawn to post-processualist subject matters and approaches. Although, it is clear that to date this potential has remained largely unexplored.
Chapter 8 – Conclusions, contributions and future work

Of challenges and opportunities

‘The truth is always something that is told, not something that is known. If there were no speaking or writing, there would be no truth about anything. There would only be what is.’ (Sontag 1963 p. 11)

Introduction

As we near the conclusion of this study, it is time to review, to summarise and to reflect upon the central research questions, practices and theories that have been expounded by the project. My purpose here is not to simply regurgitate all that has gone before this point but rather to bring together the various strands of thinking that have weaved their way through the preceding seven chapters into a coherent narrative whole. It is time to recall exactly why it was that this thesis was written in the first place and to critically assess its contributions to the answering of these questions. The chapter concludes by looking to the future and speculating upon what might lie in store both for this research project and for the wider field of Archaeological Semantic Web studies.

Chapter review

Chapter 1 introduced the central research question of this thesis. It presented a view of archaeology, as a field whose research method appears on first inspection to be deeply engaged with digital practice. Most archaeological researchers now use digital tools as a staple of their workflows (Morgan 2012 p. 1; Zubrow 2010 p. 1), and the web is the first port of call for the majority of new archaeological investigations. Archaeologists have become adept at sifting through the web’s data noise so as to find the ‘insight’ at the end of the rainbow of irrelevance. The information exchange environment allowed for by the democratisation of digital methods, which has been taking place over the last 20 years, has heralded a once in a lifetime research transformation for most archaeologists. It has changed and continues to change the way that we access our data, the scale of that data and the analyses that it can be subjected to. The digital method has become as ubiquitous to the archaeologist’s toolkit, as the pen and paper once were in the pre-digital age and,
with all likelihood, it will go on to evolve the research practice paradigm for many years to come.

And yet, once you begin to strip away the outer layers of this model, to deconstruct this digital research environment, you very quickly uncover cracks that lie just beneath its sleek veneer. The fact that archaeology is now producing more raw data than ever before is indisputable and the digital medium is the primary driver of this phenomenon (Dunn 2011 p. 97; Kansa 2011 pp. 10–11). Paradoxically however, it is becoming apparent that the potentiality of having access to more data does not necessarily equate to a better-informed archaeological community. The reality is that sizeable proportions of this total archaeological digital dataset are currently, for all intents and purposes, invisible to the researchers that would benefit from having access to them. Instead of a network of highly integrated and interdependent data points, what the online archaeological data system has become is more akin to an archipelago of independent data islands, each ignorant of the next, and the implications of this for the archaeological researcher can be wide-ranging and profound (Limp 2011 p. 271). Data invisibility can manifest itself in a number of different ways. First, while an archaeologist might be skilled in the ways of searching the web of documents, without being aware of it, they are often asking the right questions of the wrong sources. Most digital outputs of archaeological research will not appear in the results of a web search engine. The irony of the situation is that while Google and its alternatives are very good at finding information in the human-readable content of the document web, they are much less likely to uncover relevant resources in information that is intended for machine consumption.

And even if this first challenge can be overcome, then comes the related problem of what to do with the resource once you have it. The archaeological community has long since found problematic the idea that its digital data practices should be constrained by the use of discipline-wide standards (Boast & Biehl 2011; Richards 1985 p. 6; Wendrich 2011 p. 226). The effect of this decision is that archaeological projects are for the most part free to structure and serialise their data as they choose. While this gives the projects a great deal of intellectual and practical freedom, the knock-on effect is that their eventual consumers will have more work to do when extracting meaning from these resources. Chapter 1 tees up the core debate of the thesis by asking what strategies might be put in place to address the challenges of such a complex research environment.
As with all research, this thesis is surrounded by a lattice of existing and in some cases mature scholarship. The topics that it presents reference developments in the wider spheres of digital technology and the postmodern social dynamic. As such, Chapter 1 also served to situate the work within the context of the fields of first digital humanities and then digital archaeology. A five-part model for the digital humanist was presented, which divides the practice of digital humanities into separate patterns of behaviour. DHer 1 is the baseline digital humanist, DHer 2 the advanced tool user, DHer 3 the creator of new DH tools, DHer 4 the project manager, and DHer 5 the domain knowledge specialist. This model helps to position this current work in relation to other work within the DH sphere. It also provides the framework for the Archaeological Semantic Web user model, which appears later in the thesis. Chapter 1 also presented a number of models for the concepts of archaeological data and the archaeological digital data object, as well as introducing the field of digital archaeology as a related but separate intellectual space to DH.

In Chapter 2, the Semantic Web model was presented as a workable solution to the thesis’s central question. It begins by introducing two of the most important technological advances of the last century – the World Wide Web and Web 2.0 – and then explains how the idea of a Semantic Web came out of these paradigms. Gaining an understanding of this history, specifically the reasons why it was felt by Berners-Lee and others that it was necessary for the document web to evolve into a web of data, is a key component of this thesis, as it lays the essential groundwork for the rest of the work. Chapter 2 also presented the first technical topics of the thesis in the form of its primers on Linked Open Data and SPARQL, both of which serve to apply the Semantic Web’s philosophy. Lastly, Chapter 2 introduced the model of the Archaeological Semantic Web, as a knowledge domain within the overall Semantic Web, created by archaeologists to host semantically rich and interlinked digital resources for their community.

Leading on from Chapter 2’s account of the Semantic Web’s adoption by archaeologists, Chapter 3 presented a detailed overview of the ontology. While the majority of narratives on the subject of the Semantic Web present the ontology purely as a tool of Knowledge Representation, Chapter 3 bucks this trend by looking at the origins of the ontological idea within the philosophical tradition. This desire to present a more contextual picture
of the technologies that underpin the Semantic Web is seen throughout this thesis, with the rationale being that to look beyond a tool’s superficial form is to appreciate it not only as a functional entity but also as a product of history (Wartofsky 1979 pp. 177–8). This approach helped in the instance of the ontology to critically assess the design decisions taken in the building of the CIDOC CRM and its English Heritage and ARIADNE extensions, which feature prominently within the Archaeological Semantic Web implementation chapters of the thesis.

Chapters 4-6 are presented as a block. Whereas, the previous chapters sought to provide the theoretical foundations and background historical contexts for the Archaeological Semantic Web, this middle section describes the implementation of the model’s practices. linkedarc.net, which is the central technological contribution of the project, was introduced and deconstructed in Chapter 4. To emphasise the iterative and experimental nature of the system, its entire developmental history was recounted from its native RDF file system beginnings, through the intermediate MySQL hybrid triplestore and, finally, onto the Apache Jena Fuseki-based system that it has become. While Chapter 4 talked about publishing generic content to the Semantic Web, Chapter 5 specifically addressed the requirements of the archaeological community. It described the mapping of the Priniatikos Pyrgos Project dataset from its FileMaker origins to a CIDOC CRM + CRM-EH RDF output; thereby, allowing it to be published on the linkedarc.net platform. The main lesson to be taken out of Chapter 5 is that creating RDF data that conforms to the CIDOC CRM and its CRM-EH extension is an incredibly challenging undertaking. It demands a high level understanding of fundamental archaeological information structures and a good awareness of the specificities of the archaeological project in question. At the same time, the data ‘mappers’ need to be familiar with a whole range of digital techniques, which are not natural bedfellows with traditional archaeological practice.

Chapter 6 shifted the focus away from viewing the Archaeological Semantic Web as a concern of the data producer to one which showcased the model as an object of the data consumer. The chapter begins by returning to the data of the Priniatikos Pyrgos Project. It outlines the methodologies employed and results obtained by asking a number of different questions of the project’s data as it is presented on the linkedarc.net system. The objective of these examples was to show that Archaeological Semantic Web
engagement varies depending on the question being asked. There is, however, one exception to this general rule and that is that SPARQL queries are used as the data retrieval mechanism throughout. In the first Priniatikos Pyrgos case study, a SPARQL query was constructed to ask the linkedarc.net SPARQL endpoint for a description of the physical relationships that exist between the contexts of Trench II at the site. This information was then visualised using the Gephi tool, which is specifically designed to visualise these sorts of networks. In the next data-mining experiment, the loomweight assemblage of Priniatikos Pyrgos was taken as an example of a material-specific and more advanced form of Archaeological Semantic Web engagement. The PPTextiles web app presents a spatial distribution view of the find-spots of these loomweight artefacts. The distribution can be customised on the basis of the loomweight’s type or its date. An overview of the web app’s design is also provided in Chapter 6. The point of this second case study is to emphasise SPARQL’s flexibility as an interface into an Archaeological Semantic Web data source. As has been repeated several times in this thesis, the Archaeological Semantic Web in its raw state is not designed to be consumed by humans. These case studies illustrate the types of engagement that the Archaeological Semantic Web and specifically SPARQL were designed to accommodate. While the loomweight data is displayed as tabular data by the linkedarc.net web app, the PPTextiles web app, with a relatively minor expenditure of effort, is able to display the same source data using an entirely different representational approach. Chapter 6 argues that this type of methodological plasticity empowers archaeologists to engage with datasets in ways that promote the growth of new ways of thinking.

In the second half of Chapter 6, another data-mining exercise is described but in this case the source data is accessed through the British Museum’s SPARQL endpoint. The objective here was to show that the same basic principles as were followed in the Priniatikos Pyrgos examples hold for any Archaeological Semantic Web resource. The British Museum is also relevant to the Priniatikos Pyrgos RDF dataset because it too employs an extensive CIDOC CRM mapping of its collection. The reason for choosing the cuneiform digital collection and not some other category of artefact or conceptual type hosted by the British Museum was because of the range of research perspectives that this material type invites. For instance, Chapter 6’s account shows how it was possible to exploit the linked nature of Archaeological Semantic Web content by presenting a contextual narrative of the cuneiform inscriptions. First, the artefacts
containing the inscriptions themselves were retrieved and analysed. Then, the collectors associated with these artefacts were obtained, and this series of data was used to create an interactive digital timeline of the collection’s archival history. And last, the transliterations of these inscriptions were subjected to a basic level of Natural Language Processing; named entities, such as place-names and the names of individuals, were automatically extracted from the transliterations using the Stanford Named Entity Recognizer application. The lesson learnt from this last exercise was that data mined from the Archaeological Semantic Web can lead the scholar off in a number of different directions, each requiring the use of its own particular set of methods and delivering a unique set of results.

The focus and tone of the thesis shifted again as we moved into Chapter 7 and turned our attention to a consideration of the implications of the adoption of Archaeological Semantic Web practices within the field. To emphasise the fact that the use of the Archaeological Semantic Web approach involves the rejection of alternatives, the chapter began with a discussion of the FileMaker, MySQL and NoSQL systems in the context of their use as Archaeological Information Systems. The objective here was to consider how these various technologies might be employed to address the same basic challenge of promoting a more connected digital archaeology. Ultimately and unsurprisingly given the motivations and contexts for their development, each was found to fall short of the package offered by the RDF + SPARQL framework.

Chapter 7 then turned to the subject of viewing the Archaeological Semantic Web as a sociological concern. Again, the intention here was to correct the dominant trend within the field of Archaeological Semantic Web scholarship, whereby the needs and aspirations of the consuming archaeological public are sidelined in favour of presenting the model as a technological and production concern. The chapter summarised the findings of recent research into the subject, as well as presenting the results of new research carried out by this project. While this analysis indicated a number of interesting patterns in the data, the one that stood out was the disparity of views between Archaeological Information System producers and consumers on the matter of standards. Essentially, the archaeological data management cohort was seen to be in favour of standardising the way that clients gained access to their data. They supported the use of public ontologies, controlled vocabularies and data representation formats such as RDF, all of which, as we
know, are hallmarks of the Archaeological Semantic Web model. This contrasted with the consumer’s view, which rejected these models primarily on the basis that they introduced constraint and increased technical difficulty into the workflow of their research.

Chapter 7 also presented the findings of a series of workshops on the subject of the Semantic Web carried out as part of this project. The objective of these workshops was to further investigate these survey data claims. The question that the workshops posed was how tenable is the argument that the Archaeological Semantic Web would never achieve wide-scale endorsement within grass-roots archaeology on the basis that it was simply too difficult to use. While the workshops did indeed indicate that the technologies involved demanded the investment of a certain degree of time and effort on the part of the researcher, it was argued that the level of this investment was commensurate with analogous digital tools used by humanists, such as TEI. These insights were then used to formulate a model for the use of the Archaeological Semantic Web, which essentially implements the more abstract DHer framework introduced in Chapter 1.\textsuperscript{250} ASWer 1 is the standard user of the Archaeological Semantic Web. They know how to write SPARQL queries and to interpret SPARQL results. ASWer2, like DHer 2, applies their knowledge of specific digital tools, such as GIS or visualisation techniques, to Archaeological Semantic Web derived content. Lastly, ASWer 3 creates tools, such as the PPTextiles web app, that consume services of the Archaeological Semantic Web and which are intended for use by members of the archaeological community.

Given that this ASWer user model imposes certain restrictions on the use of the Archaeological Semantic Web, Chapter 7 also considered what this might mean for archaeological epistemology. It concluded that because this model requires its users to first engage in a period of formal training, this would imply that the model is more consistent with the mores of classical epistemology and not the radical tenets of some interpretations of digital epistemology. And this insight might also to some extent explain the degree of apathy that the Archaeological Semantic Web has to date been met with by significant proportions of the archaeological community, particularly those with an attraction towards more postmodernist theoretical approaches.

\textsuperscript{250} With the exception of the DHer 4 and DHer 5 user types, which were deemed to be incompatible with the view of the archaeological researcher put forth in this thesis.


**Summary of conclusions and contributions**

In the introductory chapter of this thesis McLuhan’s four Laws of Media were presented as a model on which the central arguments of this thesis could be framed. The idea was that having internalised the model, the reader could then use it as a guide with which to navigate the body of the text. Now that we are nearing the end of our study, I want to return to McLuhan’s laws and to use them to summarise the impact of the Archaeological Semantic Web from this decidedly more humanistic perspective. If you recall, the laws are described as follows:

1. Amplification: how does this tool enhance or amplify aspects of human function?
2. Obsolescence: what previous tool does its use eclipse?
3. Retrieval: does the use of this tool bring back a previously obsolete activity?
4. Reversal: when used to its fullest potential, does the tool morph into something new?

We will now consider what the asking of each of these questions tells us about the nature of the Archaeological Semantic Web when viewed as a tool of the consumer and in the process we review the contributions that this study has made to the field as a whole.

**Amplification**

There have been a number of recent studies that have looked at particular aspects of the Archaeological Semantic Web and all of these agree that the model has the potential to greatly enhance the archaeological function, as McLuhan would put it (Byrne 2008; Isaksen 2011; Nurmikko-Fuller 2015; Wright 2011). The Archaeological Semantic Web encourages openness, sharing, collaboration and the decentralisation of knowledge. It vastly increases the scale of data that any one archaeological researcher can access. It attenuates the amount of duplication of effort that has become a feature of ‘closed’ digital archaeology, as more data and services are made publically accessible, which encourages their re-use – English Heritage and the Getty’s publishing of their controlled vocabularies exemplify this spirit. The Archaeological Semantic Web’s ontological transparency opens up datasets to be consumed by a community of archaeologists, which may have no previous knowledge of the dataset – see the data mining of the British Museum’s cuneiform collection, which is structured using the CIDOC CRM, in Chapter 6 as an example. Ontological transparency also means that new data providers can learn...
from the examples set by previous data creation projects – the modelling of the Priniatikos Pyrgos dataset followed the precedents set by groups such as CLAROS, the British Museum and DBpedia – and this promotes best practice within the sector (Berrueta & Phipps 2008; Ding et al. 2010 p. 199). Re-use of existing services also tends to result in increased levels of investment being directed towards those projects that become popular, making them stronger and more relevant to contemporary practice, as a result – take the history of the CIDOC CRM’s development as a case in point. The fact that Semantic Web content is designed to be consumed by machines means that archaeologists are able to build applications that can easily access data from various different sources – take the example of the linkedarc.net web app’s Workbench feature, in which the skos:exactMatch predicate is used to find comparable records in Europeana’s and the British Museum’s datasets. While the ontological structures of these Archaeological Semantic Web datasets might differ, the underlying data representation is constant (as RDF), which means that generic tools can be built to interrogate different datasets – the LodLive and Q&D RDF browser tools, which will be described later in this chapter, are good examples of this type of generic Semantic Web data viewer.

As has been alluded to in the examples given above, the output of this research project has been twofold. On the one hand, there is the traditional doctoral thesis that you hold in your hands, and then there is the accompanying set of digital outputs, which are of equal importance. The central component of this latter group is the linkedarc.net family of solutions. I describe it as a family because it exists as both a server-side backend, in which the data of its partner projects is stored as RDF, SPARQL queries are interpreted and web services exposed, and it is a series of clients, which present particular views onto this server data. While the linkedarc.net web app is the primary human-readable representation of the linkedarc.net server data, it is not the only one. Another example, which is also described in this text, is the PPTextiles web app.251 While the range of data addressed by this second client – the loomweights of Priniatikos Pyrgos – is more limited than the ‘total’ data view presented by the linkedarc.net web app, it does, nonetheless, derive its data from the same source. The mining of the linkedarc.net server data to create visualisations of the Priniatikos Pyrgos stratigraphy using Gephi, which is described in Chapter 6, can be understood in the same terms.

251 http://pptextiles.linkedarc.net
The linkedarc.net web app acted as something of an experimental test-bed during the project. While it was necessary to ensure that it delivered a certain baseline level of functionality (it is, after all, the primary Archaeological Information System for the Priniatikos Pyrgos Project), it was also important to use it as a space within which to test out the many technical and philosophical debates that were proposed within the text. For instance, each of the data-mining exercises that are recounted in Chapter 6 are implemented in a dedicated section of the website. This is important because the majority of these experiments employ the use of interactive outputs and while screenshots of these data visualisations are often included in the text, the very nature of digital interactivity means that the sense of the experiment can only be fully appreciated when it is actually engaged with using a digital medium (Fast & Sedig 2010 p. 170). The website also allowed me to carry out other types of research, which would have been significantly more challenging were they to have been attempted using non-digital means. For instance, the Surveys section on the site was used to tally the opinions of the archaeological community on a number of matters relating to the Archaeological Semantic Web. Again, while these survey results are displayed as images in this text, the response data is also made available on the website as a raw data dump, as modelled RDF data via the SPARQL endpoint and as interactive summary visualisations.

Another significant contribution of this project has been its creation of the Priniatikos Pyrgos RDF dataset. This dataset, which is modelled primarily on the English Heritage archaeological extension of the CIDOC CRM, currently holds in the region of four million RDF triples. These represent the digital record of six years of excavation at the site and five years of post-excavation study of the materials recovered. The data is stored primarily in the linkedarc.net triplestore but it is also published as a dataset on the DataHub website. The value of making a dataset available on DataHub is not only that it will be rendered visible to more potential users, which should foster the creation of new interpretations of the data, but DataHub also acts as the primary repository on which the LOD Cloud report is prepared and so the Priniatikos Pyrgos data will soon

252 http://linkedarc.net/datamining
253 http://linkedarc.net/surveys
254 It also includes mappings to portions of other commonly used public ontologies such as Dublin Core, FOAF and RDF Schema.
255 4,259,534 RDF triples as of 11 September 2015.
256 http://datahub.io/dataset/linkedarc
become part of the official Linked Open Data Project cloud, which is an important step in realising the original objectives of the project (Bizer et al. 2009).

None of these contributions came about without the expenditure of a lot of effort, and this has not gone unnoticed in the conclusions reached. The building of the linkedarc.net server, its various clients and the creation of the Priniatikos Pyrgos dataset consumed several years of the project’s schedule. There is no doubt that producing Archaeological Semantic Web content is technically and intellectually demanding, particularly when it involves conforming to a public ontology such as the CIDOC CRM. While some automated tools, such as the ADS’s STELLAR (Wright 2011 pp. 179–186) have been created to address these sorts of challenge, I would argue that producing archaeological RDF data remains out of the reach of most small-scale archaeological projects. However, this does not mean to say that I consider myself an advocate of the decidedly pessimistic picture painted by a number of different commentators on the subject. While Isaksen (2011 p. 152) and others (Carver 2013; Kansa 2014) interpret the RDF creation problem as a potentially existential matter, I would be hopeful that projects such as ARIADNE, which seem committed to the basic principles of the Archaeological Semantic Web, would start to produce infrastructural upgrades that will allow their partner projects to port existing data onto the Archaeological Semantic Web. While ARIADNE is concerned solely with European archaeological data, I would suspect that, if successful, its model might encourage similar initiatives across the world. Of course, this centralisation model for digital archaeological data does not come without its epistemological implications. In many ways, the linkedarc.net project is an effort to fight against this tendency within the field. However, I am prepared to admit that, in the interests of pragmatism, these may be necessary concessions in order to allow the Archaeological Semantic Web to get up and running. Once this foundation is in place, then perhaps greater decentralisation agendas can be pursued.

The consumption of Archaeological Semantic Web resources also involves technical challenge but not of the same order. In recognition of these demands, the idea of the ASWer model came into being, and this is the primary non-technical contribution of the work. For the first time, it codifies the requirements that are to be met, if a researcher is to become a profitable user of the Archaeological Semantic Web. It also helps to provide

257 The LOD Cloud project is described in Chapter 2.
a more balanced view of the model as a whole. Up until now, the Archaeological Semantic Web had largely been considered as a knowledge production concern. The ASWer model serves to correct this by reminding us that a knowledge system is not tenable, if it excludes consideration of the consumer. The ASWer model, in its tripartite separation of user types, also highlights the diversity of practice that the Archaeological Semantic Web allows for.

**Obsolescence**

As with his law of amplification, McLuhan’s obsolescence law requires that we approach the Archaeological Semantic Web both as an ideal and as a practical reality. In theory, a discipline-wide adoption of Archaeological Semantic Web methods would render large extents of existing and well-established Archaeological Information System practices obsolete. To begin with, the Archaeological Semantic Web should see an end to the use of Archaeological Information Systems, which encourage the closing off of datasets to the outside world. This would mean that desktop applications such as Excel, Access, OpenOffice Calc and FileMaker would all be abandoned in favour of cloud-based systems running on backends powered by a combination of an RDF data layer and a SPARQL data interface. The same fate would await the many thousands of Archaeological Information Systems built upon MySQL data infrastructures. Either these systems would need to be extended, using RDBMS to Linked Open Data converters such as D2RQ (Abadi et al. 2007; Bizer & Seaborne 2004), or it would be necessary to map their data over to native RDF-based systems. In such a research environment, most archaeological digital data processing would move onto the digital cloud. For instance, data analytics would be carried out using services such as Plotly and IBM’s Watson Analytics.

However, in practice, even though the Archaeological Semantic Web tool should theoretically render certain aspects of Archaeological Information System use obsolete, because of certain sociological factors (unwillingness to share, privacy and legal concerns, et cetera), co-existence will be the order of the day. This means that Archaeological Semantic Web services will increasingly need to anticipate and build in functionality, which allows them to interface with these different types of Archaeological Information System.
Retrieval

As was noted in the context of McLuhan’s first law, the Archaeological Semantic Web creates the potential for a number of very positive and varied developments in the archaeological research space. However, if required, they could all be reduced to the principal idea of openness. Openness is a powerful metaphor within knowledge. Since the time of Plato’s academy, the principle that knowledge can be generated through a process of open questioning and dialogue has been championed by those of a sceptical persuasion (Vogt 2015). Science, in its post-Enlightenment form, is largely constructed on a model of open enterprise (Long 2003). In contrast to activities that are conducted in the express pursuit of financial gain, the practice of science is rewarded by the accumulation of new knowledge and secondly by the acclaim of one’s peers, and for this to happen, the agency of science needs to be open. Recent calls for increased openness of academic practice and outputs would suggest that to some extent these previous ideals might now be coming under pressure. It is plausible to suggest that academics are now less likely to release their raw data to their communities because of the pressures that are exerted on them by their institutional backers. Academics are constantly required to justify their existence in today’s more straitened times, and this means that data has become a valuable commodity with which institutional reward can be garnered (Ioannidis 2014). In this context, the practices endorsed by the Archaeological Semantic Web can be seen as a retrieval of these traditional values of openness within the academy. The other important point to make on this matter of openness is that the Archaeological Semantic Web is not simply a moral exponent of greater data openness within academia, it also provides a technical means by which to achieve this aspiration. Linked Open Data, as it is realised by RDF, and licensing schemes such as Creative Commons and Open Data Commons, allow academics to make their data publically available today.

A second retrieval of past praxis that can be encouraged by the use of the Archaeological Semantic Web tool can be viewed as decidedly less positive. The enforcement of particular ways of thinking is generally considered to be a negative phenomenon of a knowledge system, being associated with a lack of critical thinking (Micheli 2015 p. 739). As has been argued in Chapter 7, the Archaeological Semantic Web’s promotion of the use of public ontologies and controlled vocabularies has a necessarily dampening effect on the more liberal ontological tendencies of its practitioners. For instance, using the CIDOC CRM as the model for an archaeological dataset imposes a particular ontological
view onto that data. In this sense, the Archaeological Semantic Web can be seen to recall
the practices of dogmatism.

**Reversal**

The last of McLuhan’s laws cannot be applied to the Archaeological Semantic Web, as
we do not yet know what the model will transform into when it enters its mature stage of
use. However, there is nothing to stop us speculating on what this next generation of
Archaeological Information System might look like and what practices it might
encourage. Therefore, in the penultimate section of this last chapter we turn our
attention to the Archaeological Semantic Web of the future.

**The future of the Archaeological Semantic Web**

An effort has been made throughout this project to maintain a pragmatic view of the
Archaeological Semantic Web’s potential to aid the flow of archaeological information in
the 21st century. The model has been shown throughout as it currently is and not as it
possibly might become. As this thesis draws to a close, I would like to briefly correct this
rule by indulging in a degree of speculation about the types of changes that we might see
happening in this space in the coming decade. This section can be read as a form of
loose policy document for the Archaeology Semantic Web. It deliberates on what might
come to pass and where research investment should be directed.

**SPARQL changes**

In Chapter 7 we outlined a utopian view (or perhaps dystopian, depending on your
perspective) of the Archaeological Semantic Web, as a radically dynamic knowledge
environment, in which all knowledge domains are interlinked and responsive to changes
throughout a single network.258 To achieve such an outcome, even in a more limited
sense, would require a number of changes to be put into place in the design of SPARQL
data interfaces.

*Reduce data repetition using SPARQL engine linking*

One major deficiency of the Semantic Web system as it currently stands is that it does
not allow for the querying of data across multiple datasets. Currently, a resource located
in one dataset might link to resources in a second dataset but there is no simple way to

---

258 The objective of this vision should not be misunderstood as an endorsement of the types of
epistemological and sociological outcomes that it would encourage. Rather, it can be read as a
thought experiment with which to think about the natural tendencies of the Archaeological
Semantic Web and to then question whether these are of value to archaeology.
allow a consumer of that data to traverse these two datasets within a single query. This means that all of the information that is associated with the second resource cannot be accessed unless the first dataset copies these referenced RDF triples across to its own triplestore (Figure 169). The problem with this approach is that the copying action must be carried out at a point in time, and it, therefore, represents only a snapshot of that data. If the second dataset’s resources were to be updated, this change would not be reflected in the copied triples in the first dataset.

Figure 169: an example of SPARQL’s inability to query across datasets – note the inaccessible data

A solution to this would be to augment the functionality of SPARQL engines by allowing them to query across multiple RDF datasets, if required by the user’s query. The logic of this could be implemented as follows:

1. Compute all of the requested triple patterns that can be satisfied by the home RDF dataset.
2. If a triple pattern cannot be found within the home dataset, then check to see if a SPARQL endpoint exists for the missing triple’s subject URI. How would the home SPARQL engine know the address of the remote SPARQL endpoint? A very simple way of achieving this would be to assume that the endpoint exists at the following address: http://example.org/sparql in cases where the missing triple’s subject has the URI http://example.org/x/y/z. A HTTP OPTIONS request could be sent to check if a SPARQL endpoint is operational.

259 Paulheim and Hertling (2013) discuss alternative approaches to the problem of SPARQL endpoint discovery.
3. If the remote SPARQL endpoint is validated, then the home SPARQL engine could send the former a newly constructed SPARQL query that will retrieve the remainder of the requested data.

4. These checks could be implemented by all SPARQL engines and this would allow SPARQL queries to chain from one RDF triplestore to the next.

While attractive, this approach introduces a potential problem of increased latency as queries ‘jump’ from one SPARQL engine to the next. However, in practice, it is likely that most scenarios would involve only a single jump or a series of single jumps from the home dataset to a set of linked datasets. While this would need to be tested in practice, given the added benefit of this change, this added latency might be acceptable.

**Semantic Web search engines**

One of the major catalysts for the growth of the document web was the emergence of search engines, such as WebCrawler, AltaVista, Yahoo and of course, Google. For the first time, users were able to find web resources on the basis of their contents and not their URLs. To date, Semantic Web search engines have been almost entirely unknown – Swoogle is one rare exception (Ding et al. 2004; Hogan 2014). It is staggering to think that the Semantic Web has been around for such a protracted period of time without having a basic content discovery mechanism. Clearly, if the Semantic Web is to become more relevant to academic researchers, it must first address this significant shortfall (Berners-Lee 1998).

One possible way of achieving the sorts of highly specific search results that are delivered by the document web search engines would be to use SPARQL as the baseline querying interface between the Semantic Web search engine and its data sources. The first problem that would need to be overcome would be how to construct such a list of providers. There are many ways that this could be done but perhaps the simplest and most adaptable would be to use web robots to crawl the web for these datasets. These robots could use the VoID standard to identify these Semantic Web data providers (Alexander et al. 2011). The search engine could then construct a SPARQL query based on the users search request and send this to all of its known SPARQL endpoints. A variation of this general theme would be to have Semantic Web search engines that target specific knowledge domains, such as archaeology. Inevitably, the building of such systems would involve many challenges but it is work that simply must be attempted.
One of the central findings that came out of the user surveys described in Chapter 7 was that the writing of SPARQL queries is construed by potential adopters to be very challenging. SPARQL has played a leading role in this thesis’s narrative and it is the primary reason why the ASWer user model is as demanding of its participants as it is. Because of this, it is worth considering whether there might be an alternative that would be capable of delivering the flexibility of SPARQL while at the same time attenuating its barriers to entry. The solution that I propose here is to use Natural Language Processing to convert certain known natural language question types into formal SPARQL queries. A working prototype of this system is now included in the linkedarc.net web app and it functions as follows:

1. The user clicks on the new search box that is included as part of the header in all linkedarc.net web app pages. They are prompted with a set of pre-made natural language queries. For instance, ‘Where is context 1?’
2. When the search is initiated, the web app sends the natural language question to the linkedarc.net server and the server responds with a SPARQL translation of the query, if it can find a match in its database of natural language query types.
3. Once received, the web app then sends the SPARQL query to the linkedarc.net web app’s SPARQL endpoint.
4. Depending on the type of data returned, the web app then displays the results as a point or a polygon on a map (Figure 170), or as a table of results.

**Encouraging the building of more Archaeological Semantic Web tools**

Another aspect that has emerged organically out of the work undertaken by this project has been the critical role that data visualisation and other tools play in the consumption of the Archaeological Semantic Web. One of the most common complaints levelled by users against the Archaeological Semantic Web is the form of its output. Viewing RDF Turtle, RDF/XML or even an HTML representation of an RDF resource can be a confusing exercise for the Archaeological Semantic Web novice. What do the abbreviated predicate labels refer to? What is the meaning of the object data? How do you navigate through the graph of data? While it has always been understood that data representation is important to the success of the Semantic Web, for the most part little in the form of tangible tools have been released into the community. There are, however, some exceptions, which point to the types of potential gain that can be made when data is presented in an engaging and informative fashion.

![Figure 171: the LodLive tool representing the associations of the Priniatikos Pyrgos Context 1 resource](image-url)
The LodLive project is one such example of what is possible (Camarda et al. 2012). The service has a very simple working structure. By appending the URI of an RDF resource to the LodLive base URL, it will display a visualisation of the resource’s data graph. The resource’s links to other resources can be expanded and manipulated using the LodLive user interface (Figure 171). The LodLive web app retrieves this information by sending a series of SPARQL queries to the target resource’s SPARQL endpoint. As well as visualising the links between different RDF resources, LodLive monitors for the occurrence of common predicates, such as rdfs:label, foaf:depiction and geo:lat and geo:long. It uses the values associated with these to insert text labels onto resources, present a gallery of associated images (Figure 172), and to locate the resource on a map (Figure 173). This type of generic querying of Semantic Web resources illustrates the potential offered when datasets employ public ontologies such as RDF Schema, FOAF and Basic Geo. A LodLive link is now included at the end of each linkedarc.net web app record page.

Figure 172: LodLive displaying associated image resources

Figure 173: LodLive locating resources on a map

For example, http://en.lodlive.it/?http://linkedarc.net/data/la_pp/context1
Promoting the right institutional climate

Not all of the transformations required to ignite the Archaeological Semantic Web space need be technical in nature. Much work still remains to be done in convincing grassroots archaeology that the Archaeological Semantic Web is of benefit to its research. To begin with, existing Archaeological Semantic Web devotees need to do a better job of promoting the advantages that come with using the model. This point has been made before (McBride 2002 p. 419), but it is worth repeating because unless archaeologists are convinced that a method will deliver tangible benefits to their practice, they will not set out on that road of enquiry. Once researchers become interested in the model, it is then necessary to train them. With each passing year, the number of summer schools and course modules devoted to Semantic Web related topics appears to grow. This is to be welcomed, but more can always be done. Another key facet of this didactic consideration is the matter of code literacy. A prior ability to read or write code makes the internalisation of a formalised query language, such as SPARQL, substantially easier. More needs to be done towards putting in place the educational structures that serve to grow the levels of code literacy within archaeology.

A final word

While many of the arguments put forth by this thesis have been explicitly critical of aspects of the Archaeological Semantic Web, it needs to be borne in mind that this is a work of academic scholarship and for a topic to be thoroughly investigated within this context, it must be subjected to rigorous forms of questioning. If indeed the Archaeological Semantic Web is to be a realistic emancipator of archaeological digital data, then it must demonstrate that it is equipped to deal with the challenges therein. When I started out on this road of investigation, I knew nothing of the Semantic Web, of Linked Open Data, RDF or of SPARQL. However, through a process of enquiry, I have come to the opinion that despite its faults (and this thesis has shown that there are still many), the Archaeological Semantic Web represents a solid foundation on which the future of archaeological digital data can be based.

I would hope that you, the reader, leave this text with some sense of what the Archaeological Semantic Web is capable of bringing to the study of the material culture of the past. If allowed, the model truly has the potential of radically changing the

---

261 An example of this type of event is the recent Linked Pasts gathering, which was held in King’s College London in July 2015 (Wright 2015).
discipline of archaeology for the better. Projects such as Nomisma, Pelagios, the British Museum’s online collection, PeriodO and ARIADNE are no longer satisfied to wait for this change to happen, they are bringing it about for themselves. Having said that, I would be hopeful, that these larger projects do not come to dominate and to define the Archaeological Semantic Web of the future. Tim Berners-Lee and its other pioneers envisaged a decentralised space in which all data could participate, not simply that which is associated with power and influence. If nothing else, the linkedarc.net project is a gesture in that direction, an example of what can be achieved in a digital environment that champions diversity, openness and the courage to break down existing norms. It is an exhortation to build anew.

Dublin
February 2016
Appendices

Appendix A – Priniatikos Pyrgos RDF mappings using Open Refine and RDF Refine

The following accounts describe how each of the Priniatikos Pyrgos data types was mapped to an RDF implementation of the CRM-EH model. This appendix should prove of help to anyone attempting to map an existing archaeological dataset to a combination of the CIDOC CRM and its CRM-EH extension. For an introduction to the basic principles of using the RDF Refine extension, consult Chapter 5.

Mapping the image data

![Diagram of RDF classes]

Figure 174: the organisation of the RDF classes employed to map the Priniatikos Pyrgos images table

The images table has ostensibly one of the simpler data structures of any table within the Priniatikos Pyrgos project (Figure 174). It has fields for the name, description and URL address of the image. There is also a range of fields, which relate the image in question to

---

262 In the interests of legibility, the full set of predicates used in the mappings is not included in this appendix’s series of mapping schematics. For instance, the rdfs:label predicate is used to link the la_pp_ont:LA_E18_Image to the value contained in the source ‘name’ field.
other data types, such as contexts, sub-contexts, loci, pails and catalogued records. Despite this relatively straightforward structure, the images project mappings do, nonetheless, contain a number of subject root nodes in the RDF mapping skeleton. These extra root nodes are needed to accommodate the linking to other RDF resources.

As with most CRM-EH mappings, the RDF data created is heavily based upon the CRM’s event-based model. An event, conceptualised by the crmeh:EHE2010_DepictionEvent class, models the act of creating the image. Instances of this event class are then linked via the ecrm:P14_carried_out_by predicate to an actor in the form of a crmeh:EHE0077_ProjectTeamMember instance and via the ecrm:P4_has_time-span predicate to a ecrm:E52_Time-Span instance representing a timestamp for the event.
Mapping the catalogued ceramic and object data

The catalogued ceramic and object data mapping operation involved not one but two separate workflows (Figure 175). This was necessary because, as was recounted in Chapter 5, catalogued items were recorded using two separate paper record sets and this resulted in the creation of two catalogue tables within the FileMaker database. While many of the fields are shared between the two source tables, there are certain sets of fields, such as those that deal with glass objects in the catalogued objects table and those related to vessels in the catalogued ceramics table, which are unique to their table. We will, however, consider both as a group as conceptually they share many similarities and, ultimately, they are mapped to the same CRM-EH classes.

The class structure produced by the mapping of these tables is undoubtedly complex. Three key events formed the nucleus of the mapping’s ontological structure, with the first event representing the classification of the artefact in question. An instance of the
Appendix A - Priniatikos Pyrgos RDF mappings using Open Refine + RDF Refine

crmeh:EHE2002_ContextFindClassificationEvent class is linked to the catalogued pottery or object subject using the ecrm:P41i_was_classified_by predicate. This in turn contains a link via ecrm:P14_carried_out_by to an instance of a crmeh:EHE0077_ProjectTeamMember class, representing one or more archaeologists working on the project.

The second event included in the mapping represents information concerning the production of the artefact. The crmeh:EHE1002_ContextFindProductionEvent class models this type of information and it can be linked to a la_pp_ont:LA_E19_Period instance via the ecrm:P4_has_time-span predicate. As such, this series of instances can be used to model the date of the artefact’s creation; recall that this information originates from a specialist reading the artefact. And to add a further interpretive layer, ecrm:P4_has_time-span can, and generally does, contain links to more than one la_pp_ont:LA_E19_Period instance. This represents a range of possible dates for a catalogued artefact.

The third key ontological event used in the modelling of la_pp_ont:LA_E6_CataloguedObject or la_pp_ont:LA_E9_CataloguedCeramic instances is crmeh:EHE2020_ContextFindMeasurementEvent. This class represents the act of measuring some aspect of the catalogued object. Instances of this class contain links to information that describe the event in question. For example, a particular crmeh:EHE2020_ContextFindMeasurementEvent instance might account for the measuring of the artefact’s length or weight. Or it might relate to the estimation of an object’s intactness. Regardless of the type of measurement, all measurement events contain a link via the ecrm:P40_observed_dimension predicate to a crmeh:EHE0031_ContextFindMeasurement instance and this instance contains the information detailing the results of the event, i.e. the measurement values observed.
Mapping the context data

You need take only a brief look at the diagram reproduced in Figure 176 to realise that the context table mappings for the Priniatikos Pyrgos Project are extremely complex. Over 100 source fields were involved in the process and many of these required significant amounts of cleaning and reshaping in order to produce the pristine data that could be used in the mapping process. In total, it was necessary to create instances of 16 different classes for every single context record in the source project. The crmeh:EHE0007_Context class sits at the centre of the CRM-EH model. As such, it is perhaps not so surprising that the mapping is so challenging. Looking down through the list of classes, most come from the CRM-EH domain, with the only exceptions being a single class coming from the Priniatikos Pyrgos custom ontology and another from the SKOS ontology. Of the 12 CRM-EH classes, five model events,\(^{263}\) six model material or abstract concepts,\(^{264}\) and one models a human actor.\(^{265}\)

A distinction needs to be drawn between viewing the context as a conceptual entity entity and as a material object. The context concept is represented using the crmeh:EHE0007_Context class and the context object using the crmeh:EHE0008_ContextStuff class. As such, measurements of the context’s material form are associated with the crmeh:EHE0008_ContextStuff class. These measurements acts are represented using the crmeh:EHE2016_ContextStuffMeasurementEvent class, while the measurement readings themselves are represented by the crmeh:EHE0054_ContextStuffMeasurement class. The people who carried out the measurement acts are represented by the crmeh:EHE0077_ProjectTeamMember class. These crmeh:EHE0077_ProjectTeamMember resources are filled out during the Priniatikos Pyrgos personnel mapping, which is described below. The act of taking an environmental sample from the context’s material is represented in the CRM-EH model following the same basic paradigm. However, in this case, the event and its measurement are represented using the crmeh:EHE2006_ContextSamplingEvent and crmeh:EHE0018_ContextSample classes.

\(^{265}\) crmeh:EHE0077_ProjectTeamMember
Figure 176: the organisation of the RDF classes employed to map the Priniatikos Pyrgos contexts table

The crmeh:EHE1001_ContextEvent class represents the event, which created the context in the first place. This is an important class, as it contains links to other crmeh:EHE1001_ContextEvent class instances. These links define the chronological relationships between the contexts, information which is key to most archaeological site investigations (Harris 1989). Chapter 6 considers how this data might be interrogated by a consumer of linkedarc.net. The crmeh:EHE2001_ContextExcavationEvent event class collates information relating to the excavation of the context by the Priniatikos Pyrgos archaeologists. The excavators involved are modelled using the crmeh:EHE0077_ProjectTeamMember class, as would be expected, and the start and
end dates of the excavation are modelled using the crmeh:EHE0098_ContextExcavationEventTimespan class.

The final group of classes describes the finds associated with the particular context. Remember that the data used in these mappings is ultimately derived from the context sheet paper records. These sheets contained a checkbox for each of the major find types (animal none, pottery, et cetera) that were observed during the context’s excavation. These fields do not, however, allow the inclusion of the specific item counts or weights of these categories of finds. In fact, as was explained in the overview of the Priniatikos Pyrgos paper recording system, this information was instead entered into the sub-context record sheets. In that case, one might argue that the data contained on the context sheets is superfluous, as it is overridden by the more detailed information contained within the associated sub-context records. Despite this fact, a decision was made to retain as much of the digital information during the mapping process as was possible; the consumer could decide which information was relevant to their enquiry.  

The crmeh:EHE1004_ContextFindDepositionEvent class was used to model the deposition of the finds within the context. These instances then link to instances of the crmeh:EHE0009_ContextFind class, which represent the material find type itself. These find types were assigned controlled vocabulary types (in the form of SKOS instances) using the ecrm:P45_consists_of predicate.

---

266 The question of how to decide what information to include and what to leave out has huge epistemological and perhaps even ethical implications. For the most part, the linkedarc.net project attempted to map as much of the source data as was possible, even in cases, such as the scenario described here, in which the information might be interpreted as being obsolete.
Mapping the sub-context data

The sub-context table mappings for the Priniatikos Pyrgos Project are again not without their complexities and nuances. The list of source fields is reasonably long (there are 33 fields) and a large proportion of these demanded quite a bit of processing before they were ready to be mapped to RDF (Figure 177). Having said that, a number of the classes employed in the mapping should now be familiar following the descriptions of the other mappings, particularly for that of the context data. Figure 177 shows how an instance of crmeh:EHE0007_Context is created to associate the sub-context in question with its parent context entity. If you recall, the sub-contexts are administrative units, which were used at Priniatikos Pyrgos to divide up the excavation and recording of a context. They do not fit into the CRM-EH model and, as such, a custom Priniatikos Pyrgos ontology class, la_pp_ont:LA_E15_Subcontext, is used as the central subject class in the mapping. However, the sub-context table is also filled with information that is relevant to the sub-context’s parent context. You can see from the diagram reproduced in Figure 177 that the class structure linked to the la_pp_ont:LA_E15_Subcontext instance is essentially repeated for the parent crmeh:EHE0007_Context instance. In essence, this serves to
aggregate the data of a set of sub-contexts that share the same parent context and to inject this aggregation into the context’s RDF representation.

The chain of predicates and classes used to model the deposition of the sub-context finds is the same as was used to model the generic find groups in the contexts table mapping. However, while the context finds mappings did not contain information about the quantities of finds involved, the sub-context finds details are much more precise. They include sherd counts and weights divided into coarse, medium and fine groupings for pottery finds, and bag counts for all other finds. This extra information was modelled by including a link to a crmeh:EHE2020_ContextFindMeasurementEvent instance for each sub-context’s find type via the ecrm:P39i_was_measured_by predicate. The crmeh:EHE2020_ContextFindMeasurementEvent instance could then be used to map information such as the human actor involved in the measurement and the values of these measurements.

Lastly, the ecrm:P16i_was_used_for predicate was used to map the dating information applied to the pottery finds material by the project specialists. This predicate links a crmeh:EHE0009_ContextFind instance to a crmeh:EHE1005_ContextFindUseEvent representing the use that the find was put to. ecrm:P4_has_time-span was then used to link this event to one or more la_pp_ont:LA_E19_Period instances.
Mapping the locus data

Figure 178 illustrates the relative simplicity of the locus table mapping. As was explained in Chapter 5, there were no pro forma paper sheets used to record the excavated loci at Priniatikos Pyrgos. There was, however, a very rudimentary FileMaker table, which acted as a placeholder for these records in the hope that at some stage in the future they would be populated with data extracted from the excavation notebooks. The ‘geoPoly’ fields were added to the Open Refine project following the procedure outlined for the injection of geospatial data into the dataset described in Chapter 5. The name of the trench in question was derived from the name of the locus. For example, the locus ‘G1002’ was contained within the trench ‘G1000’. The RDF structure onto which this data was mapped is also straightforward. Locus entities were created as instances of the la_pp_ont:LA_E12_Locus class and this was linked to the parent trench (la_pp_ont:LA_E3_Trench) using the ecrm:P89_falls_within predicate, which is the standard CIDOC CRM way of stating that a spatial entity is located within a second spatial entity.
Mapping the pail data

The FileMaker pails information is substantially more detailed than the locus data, as pails were recorded using the same pro forma recording system used to record sub-contexts. As such, the previous explanation of the mapping of the sub-contexts data largely holds for the pails data, as well (Figure 179). The only significant difference between the two is that there is a link to a parent la_pp_ont:LA_E12_Locus instance via the ecrm:P89_falls_within predicate included in the pails data mapping in the place of the crme:HE0007_Context instance link, which is included in the sub-contexts data mapping.
Mapping the period data

The creation of a mapping for the Priniatikos Pyrgos Project’s period dataset was somewhat different to the procedure described above for the other data categories. As was discussed in Chapter 5, there was no one original paper catalogue of periods that could be queried by a team member when filling in a pro forma sheet and because of this, there was no FileMaker period table. The period dataset that was mapped to RDF in Open Refine was created as a by-product of the cleaning and vocabulary alignment carried out on the other datatables (sub-contexts, pails, catalogued entries), which contained period information. This process produced a total of 29 unique periods, each of which had a name, a start date and an end date. In keeping with the guidelines of the PeriodO project (Rabinowitz 2014), a source or reference was added to each of these period records, when applicable. Also, the relationships between the periods (for example, the Early Byzantine I and the Early Byzantine II periods are contained within the parent Early Byzantine period) were set out within the datatable.

![Source fields](image)

Figure 180: the organisation of the RDF classes employed to map the Priniatikos Pyrgos periods table

Figure 180 shows the fields contained within the period Open Refine project. The table structure is reasonably self-explanatory. It involves only a single class, ecrm:E52_Time-Span. This CRM class is used to model time and, as such, is a perfect semantic match for the meaning that the Priniatikos Pyrgos period data needed to model. Each period’s absolute start date was associated with the subject using the ecrm:P79_beginning_is_qualified_by predicate and the end date using ecrm:P80_end_is_qualified_by. When required, the ecrm:P86i_contains predicate was used to associate a parent period subject with a set of child periods. Sources were linked to the subject via the ecrm:P3_has_note predicate.
Mapping the controlled vocabulary data

As in the case of the period data, the Priniatikos Pyrgos vocabulary information was derived as a result of the cleaning of the other datatables, particularly the catalogued pottery and objects tables, whose information is acutely focused on the categorisation of material culture.

Figure 181: the organisation of the RDF classes employed to map the Priniatikos Pyrgos vocabulary table

skos:Concept was the only class utilised in this particular mapping process (Figure 181). SKOS concepts should be defined as members of a particular scheme. The skos:inScheme predicate was used to link each created skos:Concept instance to <http://linkedarc.net/vocab>, which represents the linkedarc.net controlled vocabulary scheme. Entities were each given a name using the skos:prefLabel predicate and its scope notes were linked via skos:scopeNote. The concept synonyms, which were found using Open Refine’s web service method (see Chapter 5 for more on using the Open Refine web service function), were linked to the SKOS concept using skos:exactMatch. Lastly, SKOS concepts should be published in conjunction with copyright licenses. The cc:license, cc:attributionURL and cc:attributionName predicates were all employed to signify that each of the created SKOS entities is covered by a Creative Commons Attribution 3.0 Unported (CC BY 3.0) license.
Mapping the team data

The process of mining, crowdsourcing and cleaning described in Chapter 5’s account of the compilation of the Priniatikos Pyrgos Project’s team member datatable resulted in the creation of the source field list shown in Figure 182. The actors table mapping is multi-hierarchical but is different to the other table mappings in the sense that it maps to more than one ontological model: FOAF, Schema.org and the CRM-EH. While the mapping could have been designed to create instances of the crmeh:EHE0077_ProjectTeamMember class alone, there is no reason why a mapping skeleton cannot be configured to map to different ontologies. Not all ontologies will be recognised and understood by a client and so there is added benefit in producing more than one ontological mapping.

For its part, Friend of a Friend or FOAF was one of the first Semantic Web ontologies to be widely adopted (Brickley & Miller 1999; Ding et al. 2005). It was primarily designed to model data about people and groupings of people. As such, it is concerned with describing actors and their relationship to one another. At a structural level, it is relatively simple. It includes a small number of core classes (Agent, Person, Project, Organization, Group, Document and Image) and predicates (name, title, img, depiction, et cetera). A FOAF extension adds to these lists with classes and predicates that model entities and
Appendix A - Priniatikos Pyrgos RDF mappings using Open Refine + RDF Refine

relationships within the Social Web. Each of the Priniatikos Pyrgos team table records is mapped onto an instance of the foaf:Person class with associated FOAF property mappings, such as foaf:name, foaf:givenName and foaf:familyName, also being used.

The Schema.org ontology was created by a cooperative of search engines, who wished to formalise the metadata contained within websites (Bizer et al. 2013). It is employed in this mapping process for the same reasons as was outlined for the FOAF mapping: it increases the dataset’s visibility and therefore, reusability. The Schema.org Person class is used to model the basic aspects of each of the Priniatikos Pyrgos actors. Schema.org is also multi-hierarchical and so in order to include the country information for the actor, it was necessary to also create schema:PostAddress, schema:Country and schema:GeoCoordinates instances for each team member.
Appendix B – Attitudes to digital data sharing within archaeology survey

This survey is for users of digital archaeological data services. It is an attempt to gain a snapshot of the archaeological community’s current attitude towards certain digital data sharing practices and tools. How familiar are archaeologists with the practicalities and advantages of Open Access and Open Data practice? Is Linked Data a methodology that the community might consider using at some stage in the future? Or perhaps it is already very much in use today? Is there such a thing as the Semantic Web for archaeologists and if not, why not?

This survey has been compiled by Frank Lynam as part of his doctoral research as a member of the Digital Arts and Humanities PhD programme at Trinity College Dublin. When the survey closes by the end of March 2015, all entrants will be automatically entered into a draw for a €40 Amazon gift voucher and the results of the survey will also be made available on the linkedarc.net site as Open Data under a CC Attribution-ShareAlike license.

Thanks for contributing!
Frank
Questions marked * must be answered.

The basics
How would you describe your relationship with archaeology? *

☐ I have no special interest in archaeology
☐ Archaeology is a hobby of mine
☐ I work in a cultural heritage field related to archaeology
☐ I work in contract archaeology
☐ I am an undergrad student of archaeology
☐ I am a postgrad student of archaeology
☐ I work in research archaeology as a researcher
☐ I work in research archaeology as a lecturer
☐ Other:
Appendix B – Attitudes to digital data sharing within archaeology survey

What age group do you fall into? *
- Under 16
- 16-24
- 25-34
- 35-44
- 45-54
- 55 and over

Your thoughts on digital archaeology
How familiar are you with the expression Big Data?

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>not very familiar</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

How familiar are you with Open Access practice?

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>not very familiar</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

How familiar are you with Open Data practice?

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>not very familiar</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

How familiar are you with the Semantic Web?

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>not very familiar</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix B – Attitudes to digital data sharing within archaeology survey

<table>
<thead>
<tr>
<th>Question</th>
<th>Scale</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is Big Data relevant to your archaeological interests?</td>
<td>1-5</td>
<td>not at all</td>
</tr>
<tr>
<td>Is Open Data relevant to your archaeological interests?</td>
<td>1-5</td>
<td>not at all</td>
</tr>
<tr>
<td>Is Open Access relevant to your archaeological interests?</td>
<td>1-5</td>
<td>not at all</td>
</tr>
<tr>
<td>Is the Semantic Web relevant to your archaeological interests?</td>
<td>1-5</td>
<td>not at all</td>
</tr>
</tbody>
</table>

**On digital data sharing**

Do you view sharing of data as being important to the current future success of archaeology?

<table>
<thead>
<tr>
<th>Scale</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-5</td>
<td>not very important</td>
</tr>
</tbody>
</table>

Do you currently or have you in the past shared archaeological data that you have created?

- Yes
- No
Appendix B – Attitudes to digital data sharing within archaeology survey

How happy would you be to share archaeological data that you have created?

1  2  3  4  5

not happy at all 〇 〇 〇 〇 〇 all for sharing my data

Would you like to have greater access to digital archaeological datasets?

1  2  3  4  5

not very interested 〇 〇 〇 〇 〇 this sounds fantastic

Do government regulations and/or legislation act as inhibitors to the greater sharing of digital data within archaeology?

1  2  3  4  5

not at all 〇 〇 〇 〇 〇 very much so

Do institutional work practices and regulations obstruct the greater sharing of digital data within archaeology?

1  2  3  4  5

not at all 〇 〇 〇 〇 〇 very much so

Is lack of digital skills a barrier to the greater sharing of digital data within archaeology?

1  2  3  4  5

not at all 〇 〇 〇 〇 〇 very much so

Are you as an archaeologist reluctant to share your digital data?

1  2  3  4  5

not at all 〇 〇 〇 〇 〇 very reluctant
Appendix B – Attitudes to digital data sharing within archaeology survey

Do you consider there to be a demand for greater digital data sharing within archaeology?

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

not at all  ○ ○ ○ ○ ○ huge demand
Appendix C – Semantic Web specialised user workshop

Participant survey

1. What is your profession?
   a. Data manager
   b. Coder
   c. Other (please specify)

2. What age group do you fall into?
   a. Under 16
   b. 16-24
   c. 25-34
   d. 35-44
   e. 45-54
   f. 55 and over

3. How familiar are you with the expression Big Data?
   a. 0 - not very familiar
   b. 5 - very familiar

4. How familiar are you with Open Access practice?
   a. 0 - not very familiar
   b. 5 - very familiar

5. How familiar are you with Open Data practice?
   a. 0 - not very familiar
   b. 5 - very familiar

6. How familiar are you with the Semantic Web?
   a. 0 - not very familiar
   b. 5 - very familiar

7. Have you ever used SQL before?

8. Have you ever used SPARQL before?
Exercise 1 – Did Frank Lynam excavate a lot of pottery while working at Priniatikos Pyrgos?

Objective
The site of Priniatikos Pyrgos was excavated and recorded from 2007 on using the single context system. In this system all data is tied to the central context data concept. Frank Lynam excavated at Priniatikos Pyrgos from 2007-2010. I want you to find out and represent how much (weight and sherd count) pottery he excavated during this time and then compare this to that excavated by David Govantes-Edwards.

Guidelines
I want you to use the linkedarc.net web app to access this information. You can use any of its features except for the SPARQL interface. Once you get this information, I want you to then do the same for David Govantes-Edwards. Then create a chart that compares the data for the two excavators.

Review of exercise 1

1. Did you achieve the task objectives?
2. How long did the task take?
3. If not, how far did you get?
4. Rate the difficulty of the challenge
   a. 0 – very easy
   b. 5 – very hard
5. Rate the intuitiveness of the linkedarc.net UI
   a. 0 – not very intuitive
   b. 5 – very intuitive
6. Rate the performance of the linkedarc.net UI
   a. 0 – poor
   b. 5 – very satisfactory
7. Any further comments?
Exercise 2 - Create a timeline of the Priniatikos Pyrgos excavation

Objective
I want you to create a timeline of the Priniatikos Pyrgos excavation by first asking questions of the linkedarc.net SPARQL endpoint and then visualising this data using the Timeline JS service. Ideally, I am looking for something like the timeline shown in Figure 183.

Guidelines
I am not going to give you any initial overview of the structure of the Priniatikos Pyrgos material, which is hosted on linkedarc.net, except to say that it conforms to the English Heritage extension (CRM-EH) to the CIDOC CRM (CRM). You will need to use the linkedarc.net web app and the online guides to the CRM and the CRM-EH to work this out for yourself. At a minimum you will need to retrieve each of the context names, their excavation date and their descriptions. If you have time in the end, you can supplement this by adding an image for each of the contexts.

Timeline JS is a really nice data visualisation platform and very simple to use. It allows you to create timeline visualisations for time-based data like the excavation dates for the Priniatikos Pyrgos contexts. The site includes a very easy to follow guide about how to get it up and running.
Appendix C – Semantic Web specialised user workshop

**Resources**

http://www.cidoc-crm.org/docs/CRMPrimer_v1.1.pdf

https://crmeh.wordpress.com

http://timeline.knightlab.com

**Review of exercise 2**

1. Did you achieve the task objectives (note times taken)?
   a. SPARQL query getting the excavation dates for each context?
   b. Imported data into the Timeline JS spreadsheet template in Google Docs?
   c. Display this data using Timeline JS?
   d. SPARQL query to get the context excavation notes?
   e. Incorporated this into the Timeline JS visualisation?
   f. SPARQL query to get a single image for each context?
   g. Incorporated this into the Timeline JS visualisation?

2. Rate the difficulty of the challenge
   a. 0 – very easy
   b. 5 – very hard

3. Rate the intuitiveness of the linkedarc.net SPARQL interface
   a. 0 – not very intuitive
   b. 5 – very intuitive

4. Rate the error reporting of the SPARQL interface
   a. 0 – poor
   b. 5 – very satisfactory

5. Rate the performance of the linkedarc.net SPARQL interface
   a. 0 – poor
   b. 5 – very satisfactory

6. Rate your understanding of RDF
   a. 0 – poor
   b. 5 – very high

7. Rate your understanding of SPARQL
   a. 0 – poor
   b. 5 – very high
8. Would you consider using other SPARQL services in the future?
   a. 0 – not at all
      i. Why?
   b. 5 – absolutely
      i. Why?

9. Does Linked Open Data add value to the data-mining process?
   a. 0 – not at all
      i. Why?
   b. 5 – absolutely
      i. Why?

10. Any further comments or suggestions about how linkedarc.net might be improved upon?
Appendix D – Glossary of terms

AJAX
Asynchronous JavaScript and XML. AJAX is a technology that allows web apps to request data from web servers asynchronously.

Apache Jena Fuseki
Apache Jena is a Java-based Open-source Software RDF triplestore. The Fuseki extension allows clients to query the RDF dataset using the SPARQL 1.1 formal language.

API
Application Programming Interface. A code interface (usually a set of functions) that allows an external client to interact with a computer system.

Archaeological Information System
A system designed to manage the information created by archaeological investigation. In this thesis, it is generally understood as a digital system.

Archaeological Semantic Web
A domain of knowledge relevant to archaeological investigation within the Semantic Web. This can be a broad category of knowledge as it takes in data related to material culture, history, science, the archaeological practice, art, and non-material culture.

ArchaeoML
A standard for representing archaeological information created by David Schloen of the University of Chicago.

ARIADNE
Advanced Research Infrastructures for Archaeological Dataset Networking in Europe. ARIADNE is a research institute established in 2013 to integrate the data and practices of the European archaeological sector.

ASWer
Archaeological Semantic Web user. ASWer 1, 2 and 3 are introduced in this thesis to define the types of engagement that archaeological researchers can have with the Archaeological Semantic Web.

Big Data
A term used to describe the type of data being produced in the wake of Web 2.0. Its meaning is interpreted variously by different commentators, but this thesis understands it as large quantities of unstructured data, which can only be made sense of using a machine medium.
Born digital
These are digital objects, which do not have a non-digital referent. An example of this would be an email document.

CIDOC CRM (CRM)
The CIDOC Conceptual Reference Model was created to help cultural heritage content providers structure their digital data. It has been an ISO standard since 2006.

Context
As it is understood within the Single-Context excavation and recording method, the context is a single stratigraphic event. The context might or might not have a material correlate. For example, the cutting action, which created a pit, is a negative feature. However, the fill that occupies this pit is composed of material remains.

CRM-EH
The English Heritage extension to the CIDOC CRM to model Single-Context excavation and recording data.

CRMarchaeo
The CRMarchaeo was first released in 2014. It is an extension of the CIDOC CRM used to model generic archaeological excavation and recording data. It was created under the auspices of the ARIADNE project.

CSV
Comma Separated Values. The CSV is a very simple text-based file format used to model tabular data.

Data
A low-level unit of information stored within a digital device. It is necessarily referential as it either represents an empirical observation or an interpretation.

Data mining
The act of finding patterns within a typically large source dataset.

DHer
A type of digital humanist. DHer 1, 2 and 3 are introduced in this thesis to define the range of practices of the digital humanist.

DK
DK is the abbreviation used in this text to refer to a compendium of translations of ancient Greek texts by Diels and Kranz (1951).
Appendix D – Glossary of terms

Document web
The first iteration of the web. Also known as the ‘read-only’ web or ‘Web 1.0’. The document web was typified by a very small producer base.

DOI
Digital Object Identifier. A unique label created by an official Registration Agency used to identify an internet document. The creation of a DOI is often referred to as ‘minting a DOI’.

Endpoint
An entry point into an internet system. Also known as an interface.

Gephi

Graph data
In the context of Knowledge Representation, it defines a type of data structure, which is composed of nodes and edges. Nodes are related to one another by edges.

Information
A more abstract form of knowledge than data. It is understood in this thesis as the product of data processing.

JavaScript
A programming language often used to develop client-side web apps.

JSON
JavaScript Object Notation. JSON is an easily legible format for encoding data structures. It has become a popular data exchange format on the web.

JSON-LD
A JSON-based format for encoding Linked Data.

Knowledge
The organisation of information into categories.

Knowledge Acquisition
The first stage in the knowledge creation process. Data is acquired by some means.

Knowledge Management
The output of the Knowledge Representation stage in the knowledge creation process is managed within a storage and access system.
Knowledge Representation (KR)
The data acquired in the Knowledge Acquisition stage of the knowledge creation process is represented in some form.

Knowledge Retrieval
Data, which is housed in a data repository of some sort, is accessed by an agent.

Linked Data
A foundational idea used by the Semantic Web model. Its function is to connect related data in a data network such as the web. See Graph Data.

Linked Open Data
The combination of the concepts of Linked Data and Open Data.

Locus-Pail (LP) excavation and recording
A system of excavation and recording used commonly by North American archaeological units.

Locus
In the context of the Locus-Pail system of excavation and recording, a locus defines a space that is typically delineated by architectural boundaries, most commonly walls, or an absence of cultural material.

MySQL

Notation3
An RDF serialisation format. It is very similar to RDF Turtle.

Named Entity Recognition (NER)
A Natural Language Processing technique used to automatically extract words from a block of text and to group them into categories such as the names of individuals, places and locations.

NoSQL
A database type used to host non-relational data.

Ontology
In its Knowledge Representation sense, the ontology is a formal specification of a shared conceptualisation. It allows the ontological structure of a dataset to be communicated from one agent to another.

Open Access
A set of principles, which argue that all web content be made accessible to all.
Open Data
A set of principles, which argue that all data published to the web be made accessible to all.

Open-source Software (OSS)
Computer software, which is published alongside its code. Both are made available under liberal re-use licensing schemes.

OWL
The Web Ontology Language is a meta-ontology used to author other ontologies.

Pail
In the context of the Locus-Pail system of excavation and recording, a pail subdivides a locus into administrative and to a lesser extent stratigraphic units.

Python
A general-purpose programming language.

R
An Open-source Software statistics application and programming language.

Raw data
Data, which has not been subjected to any form of secondary processing or cleaning.

RDF Turtle
An RDF serialisation format. Turtle is text-based and noteworthy for its easy legibility.

RDF/SPARQL
An RDF serialisation format. An XML-based format used as a default return type by many SPARQL engines.

RDF/XML
An RDF serialisation format. Based on XML.

RDF
Resource Description Framework is a Knowledge Representation format used to structure graph data. The basic unit of RDF is the triple.

RDFS
Resource Description Framework Schema. An RDF meta-ontology used to author other ontologies.

Semantic Web
An evolution of the web, which allows for the hosting of data instead of documents. The Semantic Web is premised upon the idea that all data be open, linkable and semantically transparent. Also known as Web 3.0.
Serialisation
A digital artefact representing a set of concepts. For example, an RDF Turtle file is a serialisation of a graph of data.

Single Context (SC) excavation and recording
A system of excavation and recording used commonly in the UK and Ireland. The central conceptual unit used by the SC method is the context. Everything else recorded in the SC method is related back to the context.

SKOS
The Simple Knowledge Organisation System is a data ontology used to model concepts. SKOS is used extensively in the archaeological and cultural heritage communities to create RDF controlled vocabularies.

Social Web
The social web is the name given to the collection of social media websites that have become popular since the mid-2000s. Also known as Web 2.0.

SPARQL
SPARQL Protocol and RDF Query Language. Pronounced ‘sparkle’, SPARQL is a formal language with which RDF data can be queried.

SQL
Structured Query Language. Pronounced ‘sequel’, SQL is a formal language with which a relational database can be queried.

String
In the context of coding, a string is a text data type.

Structured data
A block of data, whose parts can be queried. An example of a structured dataset is a relational database.

Sub-context
A custom conceptual unit used in the excavation and recording of the site of Priniatikos Pyrgos. It sub-divides the context into smaller, more manageable units.

Triple
An RDF triple is the basic data structure used by the RDF Knowledge Representation. It is composed of a subject, a predicate (or relationship) and an object.

Triplestore
A triplestore is a collection of RDF triples.
TSV
Tab Separated Values. The TSV is a very simple text-based file format used to model tabular data.

Unstructured data
Unstructured data can only be interrogated as a single unit. An example of an unstructured data type is a digital image.

URI
Uniform Resource Identifier. A unique identifier for a digital resource. The format of the URI is the same as that of a URL.

URL
Uniform Resource Locator. A sub-class of the URI. It is everything that a URI is, and it is dereferenceable on the web. This means that a web server will return data about the resource, if requested.

Web 1.0
See the entry for the Document web.

Web 2.0
See the entry for the Social Web.

Web 3.0
See the entry for the Semantic Web.

Web app
A website type, which allows for a high degree of user interactivity and usually delivers more advanced forms of functionality than simply serving web pages and images to the client. Often built on an AJAX system base.

Web of Data
See the entry for the Semantic Web.

Web service
The web service is analogous to the programming function. However, while a normal function is located within the same system as its caller, the web service is usually hosted on a web server. Web services allow web clients and servers to communicate with one another in sophisticated ways.
Bibliography


<http://crmeh.wordpress.com/>

<http://www.heritagedata.org/crmeh/crmeh_current.rdf>

——. (2014b). 5 star data - achieving the 5th star. Presented at the 13th European Networked

——. (2015a). SENESCHAL. Heritage Data. Retrieved April 9, 2015, from
<http://www.heritagedata.org/blog/about-heritage-data/seneschal/>

<http://www.heritagedata.org/blog/services/>

Datasets: Data Mapping and Extraction Via the CIDOC CRM. Christensen-Dalsgaard
B., Castelli D., Jurik B. A., & Lippincott J. (eds) Research and Advanced Technology for Digital


——. (1982). Objectivity--Explanation--Archaeology. Renfrew C., Rowlands, Michael J., &
New York.


R. S. (ed.) Hellenistic Egypt: Monarchy, Society, Economy, Culture, pp. 157–88. University of

Birger Hjørland. (2005). Empiricism, rationalism and positivism in library and information


Deployment of RDFa, microdata, and microformats on the web – a quantitative analysis.
The Semantic Web – ISWC 2013, Lecture Notes in Computer Science, pp. 17–32. Springer:
Berlin.


Hiroshima.


at the CAU Kiel: Socio-Environmental Dynamics over the Last 12,000 Years: The Creation of Landscapes IV, March 24.


Cline, B. (2012). The future of reading/thinking: epistemological construction in the age of the Kindle. communication +1, 1. DOI: 10.7275/R5KS6PH1


Europeana Project. (2014). *Definition of the Europeana Data Model v5.2.6*. Europeana.


New York: ACM. DOI: 10.1145/1807128.1807158


Lynam, F. (2013). Plan realities: reconciling the field archaeologist’s visual recording practice with the idea of the social image. *Archaeology Online*, 1/1.


Oracle. (2011). *MySQL 5.5: Storage Engine Performance Benchmark for MyISAM and InnoDB.*


