

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics
Trinity College Dublin
Ireland

A Lifecycle-Based Metadata Model for Improving Documentation, Reuse, and Generation of Declarative Mappings

A thesis submitted to the
University of Dublin, Trinity College
in fulfilment of the requirements of the degree of
Doctor of Philosophy

Sarah Alzahrani
Supervised by Prof. Declan O'Sullivan

2026

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 confirm that the work has been completed in full compliance with relevant university policies, including the Academic Integrity policy, the Trinity Policy on Good Research Practice, and the guidance on AI and GenAI in Teaching, Learning, Assessment and Research.

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.

I consent / do not consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

Signed: Sarah Alzahrani Date: 21/04/2026

ACKNOWLEDGEMENTS

I am truly grateful to my supervisor, Professor Declan O'Sullivan, for his guidance and support throughout both my Master's and PhD journey. I have learned so much from him, not only as a researcher but also as an educator. His kindness and encouragement have greatly shaped my academic path, and I am deeply thankful for his support during some of the most difficult times of my journey.

I dedicate this PhD to my father, Saeed, whom I am very proud of. He always shared with me his dream of becoming a university professor, a dream that life did not allow him to achieve. Instead, he made sure that we, his children, had every opportunity and support to reach our dreams. He is my role model, and I am especially grateful for always being by my side throughout this journey.

My deepest thanks go to my mother, Sudaa, my first teacher and my inspiration to become an educator. Her love, support, and constant prayers have guided me throughout my journey. I am truly grateful to my beloved siblings, Ghadah, Jehad, Reema, Sahar, Khalid, Hussam, and Danah, for their endless love and support. I also extend my love to my nephews and nieces, Thekra, Abdulmajeed, Lina, Yousef, and Sultan, whose joy and presence have brought so much happiness into my life. I would also like to express my heartfelt appreciation to my grandmother Sharifah and grandfather Hassan, whose prayers have always been with me.

My sincere thanks go to my dear friends, Afnan, Salwa, Sarah, Asma, Nawal and Noura, who have been like sisters to me. Their support and encouragement throughout this journey have meant so much to me.

Finally, I would like to thank my colleagues at Trinity College Dublin for their support and for creating a positive and encouraging environment during my PhD journey.

I am grateful for how this journey has shaped me into a different person and a researcher. I finally express my obedience and gratitude to God for His providence, blessings, and grace in guiding my dreams and my path.

Sarah Alzahrani
University of Dublin, Trinity College, 2026

ABSTRACT

Declarative mappings play an essential role in knowledge graph construction and data integration, enabling data transformation, alignment and interlinking of heterogeneous data sources. Despite their importance, mappings as semantic artefacts are rarely accompanied by the contextual documentation needed to understand why and how they were created. As a result, finding, assessing, maintaining and reusing these mappings becomes difficult. Existing standards such as PROV-O, DCAT, and SSSOM address aspects of provenance and alignment documentation but do not capture the full lifecycle of declarative mappings across different mapping types. This gap limits the ability to reuse, store and discover mappings in repositories, and also limits the ability of automated tools, including Large Language Models (LLMs), to generate reliable mappings without sufficient context. This thesis addresses this gap through the design and evaluation of the Lifecycle-Based Metadata Model for Declarative Mappings (LMMD). LMMD organises metadata according to five phases of the mapping lifecycle: analysis, design, development, testing, and maintenance. The model is designed to apply across mapping types including uplift mappings, ontology alignments, and entity interlinking, and can be used and extended accordingly. LMMD is then formalised as the Mapping Metadata Vocabulary (MMV), an OWL ontology aligned with established vocabularies including PROV-O, DCTERMS, and MQV. MMV is operationalised through MetaSEMAP, a web-based annotation tool that enables practitioners to document and export metadata for mapping artefacts. The thesis presents five evaluation studies. The first study is a community survey that validated the relevance of proposed metadata fields with semantic web community members. The second study assessed usability of MetaSEMAP through task-based testing. The third study examined the conceptual clarity, lifecycle coverage, and modelling quality of MMV through expert review and automated validation. The fourth study investigated whether lifecycle-based metadata supports more informed mapping reuse decisions. Finally, a use case explored whether metadata-guided prompting improves semantic correctness and completeness of LLM-generated mappings. The results demonstrate that LMMD provides a relevant,

usable, and valuable foundation for mapping documentation and reuse decisions. The findings indicate that a lifecycle-based metadata model addresses a genuine need in the knowledge graph and data integration community, with broader implications for FAIR data practices, mapping governance, and AI-assisted mapping.

Table of Contents

DECLARATION *2*

ACKNOWLEDGEMENTS *3*

ABSTRACT *4*

LIST OF FIGURES *10*

LIST OF TABLES *12*

LIST OF LISTINGS *14*

LIST OF ABBREVIATIONS *15*

Chapter 1: Introduction *17*

1.1 Motivation **17**

1.2 Research Question **20**

1.3 Research Objectives **20**

1.4 Research Methodology **22**

1.4.1 State of the Art review 22

1.4.2 Technical Approach 23

1.4.3 Evaluation Strategy 23

1.5 Thesis Contributions **25**

1.5.1 Major Contribution: Lifecycle-Based Metadata Model for Declarative Mappings (LMMD) 26

1.5.2 Minor Contribution: Metadata-Guided LLM Mapping Generation 28

1.6 Thesis Overview **28**

Chapter 2: Background *31*

2.1 From the Semantic Web to Knowledge Graph Integration **31**

2.2 Declarative Mappings in Knowledge Graph Construction **32**

2.3 Types of Declarative Mappings **34**

2.3.1 Data transformation mappings 35

2.3.2 Ontology Alignment mappings 36

2.3.3 Entity Interlinking mappings 38

2.4 The Mapping Lifecycle **41**

2.5 Metadata in the Context of Declarative Mappings **43**

2.5.1 Defining Metadata for Semantic Artefacts 44

2.5.2 Levels of Mapping Metadata 44

2.5.3 Metadata Representations in RDF 45

2.6 FAIR Principles and Semantic Artefacts **46**

2.7 Chapter Summary **47**

Chapter 3: State Of The Art *48*

3.1 Approaches to Structuring the Mapping Lifecycle 48
3.1.1 Early Lifecycle Models 49
3.1.2 Project-Centric Governance Models 53
3.1.3 FAIR-Oriented Governance Initiatives 55

3.2 Approaches to Documenting Mapping Metadata 58
3.2.1 Lifecycle-Based metadata models 58
3.2.2 Artefact-centric and Correspondence-level Models 62
3.2.3 FAIR-Oriented Metadata Frameworks 70

3.3 AI and LLM Based Mapping Generation Approaches 74

3.4 Analysis of Limitations and Open Challenges in the State of the Art 87

3.5 Chapter Summary 92

Chapter 4: Lifecycle-Based Metadata Model for Declarative Mappings (LMMD) 93

4.1 Methodology and Research Scope 93

4.2 Common Challenges in Declarative Mapping Practice 95
4.2.1 Type-Specific Documentation Practices 96
4.2.2 Artefact-Centric Documentation Limitations 96
4.2.3 Limited Lifecycle and Validation Support 96

4.3 Domain Analysis and Intended Use Cases 97
4.3.1 Stakeholders and Roles 98
4.3.2 Mapping Lifecycle Use Case Scenarios 98
4.3.3 Mapping Lifecycle Challenge and Design Motivation 100
4.3.4 Conceptual Domain Model and Usage Overview 102

4.4 Requirements Derived From Existing Approaches, Use Case Scenarios and FAIR Principles 104

4.5 Design of the Lifecycle-Based Metadata Model 106
4.5.1 Design Features and Requirement Traceability 106
4.5.2 Lifecycle Phases and Associated Metadata 108
4.5.3 Competency Questions as Design Validation Instrument 110

4.6 Ontology Implementation and Tool Operationalisation 111
4.6.1 Ontology Engineering Approach 112
4.6.2 Mapping the Conceptual Model to OWL (MMV) 113
4.6.3 Vocabulary Reuse and New MMV Terms 116
4.6.4 Illustrative Example 120

4.7 MetaSEMAP Design and Implementation 122
4.7.1 Purpose and role of MetaSEMAP 122
4.7.2 Core annotation workflow 124
4.7.3 Supported metadata representations 127

4.8 Evolution of the Model Based on User and Expert Feedback 133
4.8.1 Model Development and Initial Assumptions 133
4.8.2 Evaluation Feedback Refinements 134

4.9 Chapter Summary 135

Chapter 5: Evaluation 137

5.1 Overview of Evaluation Strategy 137

5.2 Evaluation instruments 140

5.3 Study 1: Community Survey on Metadata Requirements 141

5.3.1 Motivation 141
5.3.2 Hypothesis 142
5.3.3 Methodology 143
5.3.4 Evaluation instruments 143
5.3.5 Experiment Structure 144
5.3.6 Participants and Recruitment 145
5.3.7 Results 146
5.3.7.1 Initial Round Results 146
5.3.7.2 Follow-up Round Results 153
5.3.8 Discussion 154
5.3.9 Experiment Summary 156

5.4 Study 2: Usability Evaluation of MetaSEMAP 157

5.4.1 Motivation 157
5.4.2 Hypothesis 158
5.4.3 Methodology 158
5.4.4 Evaluation instruments 160
5.4.5 Experiment Structure 161
5.4.6 Participants and Recruitment 162
5.4.7 Results 163
5.4.8 Discussion 172
5.4.9 Experiment Summary 173

5.5 Study 3: Ontology Evaluation 174

5.5.1 Motivation 174
5.5.2 Hypothesis 175
5.5.3 Methodology 176
5.5.4 Evaluation Instruments 177
5.5.5 Experiment Structure 178
5.5.6 Participants and Recruitment 178
5.5.7 Results 179
5.5.8 Discussion 184
5.5.9 Experiment Summary 186

5.6 Study 4: Mapping Reuse Experiment Based on Metadata 187

5.6.1 Motivation 187
5.6.2 Hypothesis 188
5.6.3 Methodology 188
5.6.4 Evaluation Instruments 190
5.6.5 Experiment Structure 191
5.6.6 Participants and Recruitment 192
5.6.7 Results 192
5.6.8 Discussion 198
5.6.9 Experiment Summary 200

5.7 Chapter Summary 201

Chapter 6: Use Case - Metadata-Guided LLM Mapping Generation 203

6.1 Motivation 203
6.2 Hypothesis 204

6.3 Methodology 205
6.4 Experiment Structure 208
6.5 Results 211
6.5.1 Discussion 218
6.5.2 Experiment Summary 221

Chapter 7: Conclusion and Future Work *222*

7.1 Research objectives **222**

7.2 Research Contributions **223**

7.3 Impact and Uptake **224**

7.4 Future Work **225**

7.5 Final Remarks **226**

References *227*

Appendices *234*

Appendix A : Full Metadata Field Specifications **235**

Appendix B: CQs, MMV Properties and SPARQL Query Templates **241**

Appendix C: MMV new classes and properties tables **243**

Appendix D: Study 1 Materials **245**

Appendix E: Study 2 Materials **261**

Appendix F: Study 3 Materials **267**

Appendix G: Study 4 Materials **277**

Appendix H: Use Case: Metadata-Guided LLM Mapping Generation (Chapter 6) Materials **298**

Appendix I: Published papers **320**

LIST OF FIGURES

Figure 1.1 Overview of the main declarative mapping types 17
Figure 1.2 Overview of the evaluation strategy and its relationship to research objectives. 25
Figure 2.1 Organisation of basic correspondence patterns [27] 34
Figure 2.2 Example of an RML mapping specification illustrating the structure of a Triples Map [30]. 36
Figure 2.3 Example of an ontology alignment correspondence expressed using the Alignment API format [31]. 38
Figure 2.4 Example of a Silk link specification for entity interlinking (Volz et al., 2009). 39
Figure 2.5 General software development lifecycle phases[38] 42
Figure 3.1 OISIN ontology mapping lifecycle represented as a UML activity diagram [38] 50
Figure 3.2 OM2R ontology mapping lifecycle illustrating six phases and parallel metadata creation 52
Figure 3.3 Project-centric ontology mapping lifecycle proposed by Debruyne et al.[5] 55
Figure 3.4 FAIR-IMPACT Mapping Practice Framework illustrating five phases and stakeholder involvement[9], [40] 57
Figure 3.5 Mapping generation prompt used by Hofer et al. 76
Figure 3.6 Syntax repair prompt used by Hofer et al. 76
Figure 3.7 Zero-shot prompt configuration used in MYAM+R by Schmidt et al. 77
Figure 3.8 Simple task prompt used by Frey et al.[59] 78
Figure 3.9 Most contextually rich task prompt (T5) used by Frey et al. 78
Figure 3.10 Prompt template and example used by Babaei Giglou et al. 80
Figure 3.11 Simple and most contextually rich prompt variants used by Jimenez Ruiz et al. 81
Figure 3.12 Prompt template used by Taboada et al. in MILA for ontology concept matching 81
Figure 3.13 Example prompt used by Emonet et al. 83
Figure 3.14 Simple zero-shot prompt used by Ihtassine et al. in NL2LS 84
Figure 3.15 Refined prompt used by Ihtassine et al. in NL2LS 84
Figure 4.1 Overview of the Lifecycle-Based Metadata Model for Declarative Mappings (LMMD) 95
Figure 4.2 Summary of use case scenarios and the roles involved 100
Figure 4.3 Lifecycle phases for declarative mapping projects 102
Figure 4.4 Conceptual UML domain model for lifecycle-based documentation of declarative mapping projects. 103
Figure 4.5 Overview of the MMV ontology 116
Figure 4.6 MetaSEMAP main interface 123
Figure 4.7 MetaSEMAP annotation workflow for lifecycle-based mapping metadata 124
Figure 4.8 MetaSEMAP metadata annotation interface 125
Figure 4.9 Tooltip example: metadata field explanations 126

Figure 4.10 MetaSEMAP review screen displaying a summary of the entered metadata fields 126
Figure 4.11 MetaSEMAP post-annotation output screen 127
Figure 5.1 Overview of the four evaluation studies conducted to assess LMMD 137
Figure 5.2 Evaluation instruments and measures across four studies 140
Figure 6.1 Web-based interface used to run the LLM mapping generation experiment 208

LIST OF TABLES

Table 2.1 Comparative overview of declarative mapping types across key characteristics 40
Table 2.2 Examples of correspondence patterns across the three declarative mapping types 41
Table 3.1 Key metadata fields captured by OM2R 60
Table 3.2 Key metadata fields captured by the project-centric governance model 62
Table 3.3 Key metadata fields in SSSOM. 65
Table 3.4 Key metadata fields in the Alignment API format and EDOAL extension [34], [35]. 66
Table 3.5 Key metadata fields captured by RML approaches 69
Table 3.6 Overview of prompt context and output annotation across LLM-based mapping generation approaches 86
Table 3.7 Comparative overview of existing approaches and LMMD across key dimensions 91
Table 4.1 Requirements for the Lifecycle-Based Metadata Model for Declarative Mappings (LMMD) 105
Table 4.2 Summary of how the design requirements of LMMD align with the FAIR principles 106
Table 4.3 Traceability between LMMD design requirements and design features 108
Table 4.4 Metadata fields captured across LMMD lifecycle phases, ... 110
Table 4.5 Representative competency questions used to guide the design of LMMD 111
Table 5.1 Participant profiles across evaluation studies 139
Table 5.2 Participant breakdown by round and experience level 146
Table 5.3 Ratings results of the metadata fields relevance to mapping discovery task by Experts 147
Table 5.4 Ratings results of the metadata fields relevance to mapping discovery task by Intermediates 148
Table 5.5 Rating results of the metadata fields relevance to mapping quality assessment task by Experts 150
Table 5.6 Ratings results of the metadata fields relevance to mapping quality assessment task by intermediates 151
Table 5.7 Phase-level mean ratings and standard deviations for the initial round by experience group 152
Table 5.8 Phase-level mean ratings and standard deviations for the follow-up round by experience group 153
Table 5.9 Summary of the three scenarios 162
Table 5.10 PSSUQ scores per participant 164
Table 5.11 PSSUQ subscale means and standard deviations (n = 46). 165
Table 5.12 Per-item means for the 14-item PSSUQ/SUS instrument (n = 46). 166
Table 5.13 Task completion times by scenario 167
Table 5.14 Themes identified from qualitative feedback (n = 31). 171
Table 5.15 Quantitative scores for the expert evaluation 179

Table 5.16 Q7 Modifications 180
Table 5.17 Q8 Documentation 181
Table 5.18 Purpose Vocabulary response 182
Table 5.19 Purpose category rankings per expert 183
Table 5.20 Expert open-ended responses on additional purpose categories 183
Table 5.21 Expert open-ended responses on purpose vocabulary usage 184
Table 5.22 Scenario 1(Interlinking Rule) - Participant responses 193
Table 5.23 Scenario 2 (RML Uplift) - Participant responses 193
Table 5.24 Scenario 3 (Ontology Alignment) - Participant responses 194
Table 5.25 Qualitative feedback themes 197
Table 6.1 Evaluation criteria for S1A and S1B 211
Table 6.2 S1A results (Correctness / Completeness) 212
Table 6.3S1B results (Correctness / Completeness) 213
Table 6.4 Evaluation criteria for S2A and S2B 214
Table 6.5 S2A results (Correctness / Completeness) 214
Table 6.6 S2B results (Correctness / Completeness) 215
Table 6.7 Evaluation criteria for S3A and S3B 216
Table 6.8 S3A results (Correctness / Completeness) 217
Table 6.9 S3B results (Correctness / Completeness) 218

LIST OF LISTINGS

Listing 4.1 Example instantiation of a lifecycle-based mapping project using MMV 120
Listing 4.2 RDF-star annotation for an uplift mapping using initial metag: vocabulary 129
Listing 4.3 Named Graph annotation for an uplift mapping 129
Listing 4.4 RDF-star annotation for an ontology alignment 130
Listing 4.5 Named Graph annotation for an ontology alignment 131
Listing 4.6 RDF-star annotation for an interlinking specification 132
Listing 4.7 Named Graph annotation for an interlinking specification 132
Listing 6.1 MMV Baseline prompt structure 206
Listing 6.2 Guided prompt structure (with lifecycle metadata) 206

LIST OF ABBREVIATIONS

Abbreviations	Definition
AI	Artificial Intelligence
API	Application Programming Interface
CSV	Comma-Separated Values
DCAT	Data Catalog Vocabulary
DCMI	Dublin Core Metadata Initiative
DCTERMS	Dublin Core Metadata Terms
DL	Description Logic
DQV	Data Quality Vocabulary
EDOAL	Expressive and Declarative Ontology Alignment Language
EOSC	European Open Science Cloud
FAIR	Findable, Accessible, Interoperable, Reusable
FOAF	Friend of a Friend
GDPR	General Data Protection Regulation
GPT	Generative Pre-trained Transformer
IRI	Internationalised Resource Identifier
ISWC	International Semantic Web Conference
JSON	JavaScript Object Notation
KG	Knowledge Graph
LIMES	Link Discovery Framework for Metric Spaces
LMMD	Lifecycle-Based Metadata Model for Declarative Mappings
LLM	Large Language Model
LOT	Linked Open Terms
MMV	Mapping Metadata Vocabulary
MQV	Mapping Quality Vocabulary
OAEI	Ontology Alignment Evaluation Initiative
OISIN	Ontology Integration and Interoperability
OWL	Web Ontology Language
PROV-O	Provenance Ontology
PSSUQ	Post-Study System Usability Questionnaire
R2RML	RDB to RDF Mapping Language
RDA	Research Data Alliance
RDF	Resource Description Framework

RML	RDF Mapping Language
SHACL	Shapes Constraint Language
SKOS	Simple Knowledge Organisation System
SPARQL	SPARQL Protocol and RDF Query Language
SSSOM	Simple Standard for Sharing Ontological Mappings
SUS	System Usability Scale
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
UML	Unified Modelling Language
WIDOCO	Wizard for Documenting Ontologies
XML	Extensible Markup Language
YARRRML	Yet Another RDF Rules Mapping Language

Chapter 1: Introduction

1.1 Motivation

Declarative mappings play an essential role in knowledge graph construction and data integration, enabling data transformation, alignment and interlinking of heterogeneous data sources. They are used across application domains to transform tabular and semi-structured data into RDF, align ontologies, and interlink related entities across datasets, Figure 1.1 provides an illustration of these mapping processes [1], [2], [3]. Despite their importance, the process of developing declarative mappings remains challenging. Creating a mapping often requires resources, significant time and expertise, and even small changes in source data, modelling decisions, or organisational requirements can lead to extensive redevelopment efforts. Even when created, mappings are often lack the contextual information needed to understand how and why they were created, making them difficult to maintain or reuse [4], [5], [6].



Figure 1.1 Overview of the main declarative mapping types

One of the underlying issues is that mapping development is a lifecycle-driven activity. It involves analysis of data and requirements related to the application domain, design decisions about semantics and modelling choices, implementation of the actual mappings, testing against real data, and long-term maintenance as systems evolve [2], [7]. However, most mapping artefacts only represent the final stage of this lifecycle. The contextual knowledge that shapes a mapping, its purpose, constraints, assumptions, quality requirement, validation results, or maintenance history, typically remains undocumented or is captured informally in ways that are not in a machine-readable format. As a result, mapping files

frequently appear opaque to others, and knowledge engineers struggle to assess whether an existing mapping is suitable for a new task or reuse. Studies on mapping quality, reuse of mappings, and alignment evaluation all point to missing documentation as an obstacle to interoperability [8], [9].

A number of metadata and provenance standards seek to improve transparency in data transformation processes, including Dublin Core[10], PROV-O [11], DCAT [12], and vocabularies for assessing data and mapping quality such as DQV [13] and MQV [8]. However, these vocabularies, while valuable, are not designed to capture the full lifecycle of declarative mappings. Some initiatives focus on ontology alignment metadata, such as SSSOM [14], but these approaches address only a subset of mapping activities and do not generalise across uplift or interlinking tasks. Provenance generation techniques for RML mappings [15] can record execution traces but do not document conceptual design decisions or lifecycle rationale. Consequently, metadata practices in the mapping community remain fragmented, limited in scope, and insufficient for supporting long-term understanding or reuse.

This limitation is evident in established initiatives such as the Ontology Alignment Evaluation Initiative (OAEI)[1] [16], where mapping artefacts are shared primarily as standalone alignment files accompanied by aggregate evaluation scores, while detailed metadata about project purpose, modelling assumptions, configuration choices, and lifecycle evolution is provided only in narrative form or omitted altogether, and is not represented in a structured, machine-interpretable manner.

At the same time, international efforts such as FAIR-IMPACT[2] have emphasised the need for improved documentation and governance of semantic artefacts, including mappings and crosswalks used in data publication workflows [9]. These initiatives highlight that mappings, like datasets, require structured metadata to be findable, interpretable, and reusable. However, the ecosystem still lacks a unified

[1] https://oaei.ontologymatching.org/
[2] https://fair-impact.eu/

solution capable of representing mapping context across different mapping types and lifecycle stages. Related community efforts are also emerging to address metadata challenges more broadly, such as the Research Data Alliance (RDA) Metadata Interest Group[3] [17], which coordinates metadata principles, element sets, and interoperability practices across research domains. Building on such foundations, the RDA FAIR Mappings Working Group[4] [18] focuses more specifically on developing shared mapping guidelines and exchange models, highlighting both the urgency of the problem and the need for structured, machine-readable metadata for mappings.

Alongside these challenges, emerging research on Large Language Models (LLMs) has shown that they can assist with mapping-related tasks, such as generating transformation rules or suggesting schema correspondences [19], [20]. However, LLM-generated mappings often suffer from missing constraints, incomplete logic, or inconsistencies, in part because the models operate without access to the lifecycle context that typically informs human mapping decisions. While LLMs show potential to support mapping development, the extent to which structured metadata can guide LLMs toward producing more accurate, complete, and reusable mappings remains largely unexplored.

The absence of a lifecycle-based metadata model for declarative mappings, the lack of unified metadata approaches across mapping types, and the limited support offered by existing metadata standards reveal a significant gap in the state of the art. It is further reinforced by the growing need to understand how metadata and semantic context can inform AI-assisted mapping workflows. Addressing this gap motivates the development of a comprehensive, lifecycle-aware metadata model that can document mapping projects, support informed mapping reuse decisions, and provide context to enhance automated mapping generation. This motivation forms the basis for the research presented in this thesis.

[3] https://www.rd-alliance.org/groups/metadata-ig/activity/
[4] https://www.rd-alliance.org/groups/fair-mappings-wg/activity/

1.2 Research Question

The research question investigated in this thesis is:

To what extent can a lifecycle-based metadata model improve the documentation, reuse, and generation of declarative mappings?

The research focuses on deployment within knowledge graph/linked data pipelines. The terms used in this research question are defined as follows:

- *Lifecycle-based metadata model* refers to a structured model that organises metadata according to the phases of the mapping lifecycle (analysis, design, development, testing, and maintenance) and integrates these phases to support consistent documentation.
- *Declarative mappings* are specifications that describe how to transform, align, or interlink data from a source into a target representation. In this thesis, the term is used in a broader sense to encompass three main types: data transformation mappings (uplift), ontology alignment mappings, and entity interlinking specifications.
- *Documentation* refers to the process of generating statements about a declarative mapping, e.g. purpose, and information about its lifecycle.
- *Reuse* refers to the process of understanding, assessing, and applying existing mappings to new but related mapping development tasks
- *Generation* refers to the process of creating new declarative mappings, potentially supported by metadata-driven guidance or automation.

1.3 Research Objectives

To address the research question, the following research objectives were derived:

RO1: Identify and validate comprehensive metadata requirements relevant to different mapping types and phases of the mapping lifecycle.

a) Review existing mapping metadata standards, mapping practices, and documentation approaches.
b) Analyse gaps and overlaps to establish requirements for a Lifecycle-Based Metadata model for Declarative Mappings.
c) Evaluate the relevance and completeness of mapping metadata requirements through a community survey.

RO2: Design a lifecycle-based metadata model for declarative mappings, implement and evaluate it as an ontology that captures the identified requirements from RO1.
a) Define core concepts, properties, and relationships across phases of a declarative mapping lifecycle in order to create the lifecycle-based metadata model.
b) Design and implement the corresponding ontology, including alignment with existing vocabularies to promote interoperability and reuse.
c) Evaluate the ontology using expert review, validation tools, competency questions, and real-world mapping scenarios.

RO3: Develop a tool (subsequently named the MetaSEMAP tool) to demonstrate the proposed Lifecycle-Based Metadata model for Declarative Mappings and how it could be operationalised.

RO4: Evaluate the lifecycle-based metadata model for declarative mappings.
a) Assess the usability and user satisfaction of the metadata model and the tool through task-based studies.
b) Investigate whether the lifecycle-based metadata model improves mapping reuse decisions through a mapping reuse experiment.
c) Analyse quantitative and qualitative feedback to refine both the ontology and the tool.

RO5: Examine the potential impact of lifecycle-based mapping metadata on automated declarative mapping generation using Large Language Models (LLMs).

a) Compare mapping outputs generated with and without guided prompts using the lifecycle-based mapping metadata model
b) Evaluate the accuracy and completeness of LLM-generated mappings.

1.4 Research Methodology

This research adopts a design-science and empirical evaluation methodology that integrates conceptual modelling, technical development, and systematic validation. The methodology is structured into three main components: a state-of-the-art review to establish the research context and identify existing gaps, a technical approach detailing the design and implementation of the proposed solution, and an evaluation strategy that evaluates the approach in terms of relevance, usability, and effectiveness.

1.4.1 State of the Art review

The first phase of the research involved a review of existing work on declarative mappings, metadata models, and lifecycle management. The review focused on three key themes: (1) current practices and challenges in mapping documentation and reuse; (2) metadata and provenance ontologies that provide structures for describing mapping activities; and (3) emerging research that explores how Large Language Models (LLMs) can support automatic mapping generation.

The analysis revealed several limitations in the current landscape. Existing metadata models often concentrate on a single mapping type or lifecycle phase, leading to weak and limited interoperability. Few solutions exist that enable users to capture and reuse lifecycle-based metadata in a structured, machine-readable format. Moreover, the impact of metadata on reuse of mappings and on the outputs of AI-generated mappings remains largely unexplored. These findings motivated the design of a lifecycle-based metadata for declarative mappings that could address these gaps by supporting consistent documentation, promoting reuse, and guiding mapping generation through a metadata model.

1.4.2 Technical Approach

Building on the insights gained from the state-of-the-art review, the technical approach followed a design process that combined ontology development and tool implementation. The first stage involved the design of the Lifecycle-Based Metadata Model for Declarative Mappings, which formalises metadata requirements across the five phases of the mapping lifecycle, analysis, design, development, testing, and maintenance. Then, an ontology was developed using ontology-engineering principles, including the formulation of competency questions, reuse of established vocabularies such as PROV-O, DCTERMS, and MQV, and refinement through expert feedback to ensure conceptual accuracy and interoperability.

The second stage focused on the development of the MetaSEMAP tool, a web-based system that applies the ontology in practice by enabling users to annotate and export declarative mappings enriched with lifecycle metadata. The tool supports metadata export in both RDF-star and Named Graph formats, demonstrating how structured metadata can be applied in practice to improve mapping documentation.

Finally, the solution was extended to explore how metadata-guided prompts could be used with LLMs to examine whether lifecycle context improves the accuracy, completeness, and reusability of automatically generated mappings. These activities deliver the Lifecycle-Based Metadata model for Declarative Mappings, addressing Research Objectives RO1 to RO3.

1.4.3 Evaluation Strategy

The final phase of the research focused on evaluating the proposed model through a series of complementary studies examining metadata relevance, tool usability, ontology design, and reuse decisions, together with a use case exploring AI-assisted mapping generation.These studies provide an assessment of how the

lifecycle-based metadata model supports documentation, reuse, and automated mapping creation.

The first study, a community survey, gathered feedback from mapping practitioners to assess the perceived relevance, clarity, and usefulness of the proposed mapping metadata fields. The findings were used to validate and refine the metadata requirements and the lifecycle-based mapping metadata model design, in support of addressing RO1.

The second study, a usability evaluation of the MetaSEMAP tool, assessed the tool's effectiveness, efficiency, and user satisfaction following a usability framework and using the Post-Study System Usability Questionnaire (PSSUQ). Participants completed annotation tasks across multiple scenarios, and both quantitative metrics and qualitative feedback informed improvements to the tool design. This study was in support of addressing RO4.

The third study, an ontology evaluation, examined the conceptual clarity, lifecycle coverage, modelling structure, and vocabulary reuse of MMV. The evaluation used expert review and automated validation using OOPS! and FOOPS!, in support of RO2.

The fourth study, a mapping reuse experiment, investigated whether lifecycle-based metadata supports more informed and accurate reuse decisions. Participants were asked to decide whether mappings should be reused based on their associated metadata, with outcomes measured through decision accuracy, confidence, and qualitative feedback. This study evaluated the practical usefulness of metadata to support mapping reuse process and contributed to addressing RO4.

A use case, presented as a separate chapter, explored metadata-guided LLM mapping generation by comparing unguided and metadata-guided prompting to measure the impact of the lifecycle-based mapping metadata model on the semantic correctness and completeness of automatically generated mappings. This

use case explored the potential of using metadata about the mapping lifecycle to enhance AI-assisted mapping generation, addressing RO5.

Figure 1.2 provides an overview of the evaluation strategy adopted in this thesis and illustrates how the individual evaluation studies collectively support the research objectives defined in Section 1.2.

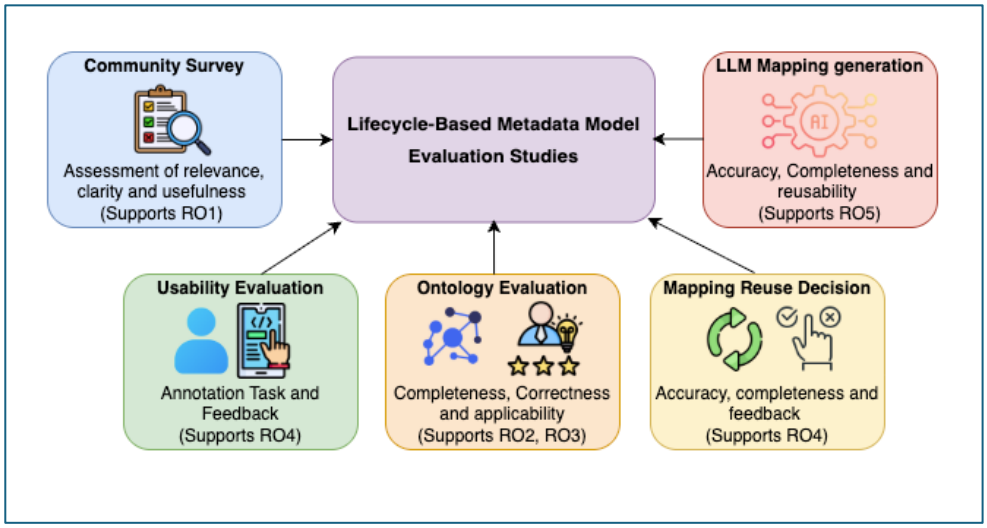


Figure 1.2 Overview of the evaluation strategy and its relationship to research objectives.

Collectively, these five studies provide quantitative and qualitative evidence of the value of lifecycle-based metadata model in improving the documentation, usability, reuse, and automation of declarative mappings.

1.5 Thesis Contributions

The research described in this thesis resulted in one major contribution and one minor contribution. The major contribution is the design and development of the Lifecycle-Based Metadata Model for Declarative Mappings, which provides a unified, lifecycle-aware foundation for documenting, maintaining, and reusing declarative mappings. This contribution is realised through its formalisation as the Mapping Metadata Vocabulary (MMV) Ontology and its demonstration implementation within the MetaSEMAP tool. The thesis also presents evaluation

through a community survey, usability testing, ontology evaluation and a mapping reuse decision experiment that demonstrates the model's relevance, completeness, and practical value.

This contribution is expected to support ongoing initiatives concerned with the governance, documentation, and reuse of mappings in knowledge graph and data integration infrastructures. In particular, it aligns with the goals of initiatives such as FAIR-IMPACT[2] and the RDA FAIR Mappings Working Group[4] by providing a structured, machine-readable way to represent mapping context, lifecycle information, and quality-related metadata.

By making such information explicit and reusable across mapping types, the model is anticipated to complement existing efforts focused on FAIR Digital Objects for mappings, crosswalk documentation, and mapping exchange models. The project-centric and lifecycle-based perspective may also benefit evaluation-driven initiatives such as the Ontology Alignment Evaluation Initiative (OAEI)[1] by enabling richer, standardised metadata about alignment context, assumptions, and evolution beyond standalone alignment files and aggregate scores.

The minor contribution is a study on metadata-guided mapping generation using Large Language Models (LLMs), which investigates how structured lifecycle metadata can enhance the semantic correctness and completeness of automatically generated mappings. This work is expected to inform future research and tool development on human AI collaboration in mapping engineering, by showing how structured lifecycle metadata can serve as lightweight guidance for producing more reliable LLM-assisted mappings.

1.5.1 Major Contribution: Lifecycle-Based Metadata Model for Declarative Mappings (LMMD)

The major contribution of this thesis is the design and development of a Lifecycle-Based Metadata Model that captures the metadata needs of declarative mappings across the five phases of the mapping lifecycle: analysis, design, development, testing, and maintenance. Associated publications include:

- **Alzahrani, S., & O'Sullivan, D. (2026).** ***MMV: A Lifecycle Based Ontology for Declarative Mapping projects,*** **In Proceedings of the IEEE 20th International Conference on Semantic Computing (IEEE ICSC 2026) -** *(In press)***.** *This paper presents MMV (Mapping Metadata Vocabulary), a lifecycle-based ontology designed to represent metadata for declarative mapping projects across different mapping types. The paper describes the design of MMV, illustrates its instantiation, and outlines a planned expert evaluation.*
- **Alzahrani, S., & O'Sullivan, D. (2025).** ***MetaSEMAP: Usability Evaluation for Metadata Annotation Across the Mapping Lifecycle.*** **In Proceedings of the International Workshop on Users and Knowledge Graphs (UKG 2025), co-located with the SEMANTiCS Conference**. *This paper reports a user study evaluating the usability and learnability of structured lifecycle-based metadata annotation using the MetaSEMAP tool.*
- **Alzahrani, S., & O'Sullivan, D. (2024).** ***MetaMap: Facilitating Mapping Reuse and Lifecycle Documentation.*** **In Proceedings of the IEEE 18th International Conference on Semantic Computing (IEEE ICSC 2024).** *This paper introduces an early prototype of MetaSEMAP supporting mapping reuse and lifecycle documentation, motivating the need for structured mapping metadata.*
- **Alzahrani, S., & O'Sullivan, D. (2023).** ***What Metadata Is Needed for Semantic and Data Mappings?*** **In Proceedings of the IEEE 17th International Conference on Semantic Computing (IEEE ICSC 2023).** *This paper presents a community survey identifying key metadata requirements for semantic and data mappings, forming the empirical foundation for the Mapping Metadata vocabulary (MMV) design.*
- **Alzahrani, S., & O'Sullivan, D. (2022).** ***Towards a Unified Metadata Model for Semantic and Data Mappings.*** **In Proceedings of the Ontology Matching Workshop (OM 2022), co-located with the International Semantic Web Conference (ISWC 2022).** *This paper explores the feasibility of a unified metadata model for mappings, providing early conceptual insights that informed the lifecycle-based model adopted in MMV.*

1.5.2 Minor Contribution: Metadata-Guided LLM Mapping Generation

The minor contribution of this thesis is an exploratory study on metadata-guided mapping generation using Large Language Models (LLMs). This work investigates the potential of lifecycle-based metadata to support automated or semi-automated mapping generation. Associated publication:

- **Alzahrani, S., & O'Sullivan, D. (2025).** ***Guiding LLM-Generated Mappings with Lifecycle-Based Metadata: An Early Evaluation.*** **In Proceedings of the KG-NeSy Workshop, co-located with the SEMANTiCS Conference 2025.** *This paper investigates whether providing lifecycle-based metadata as part of LLM prompts improves the semantic correctness and completeness of declarative mappings generated by large language models.*

The full versions of these *publications* are included in Appendix I.

1.6 Thesis Overview

Chapter 2: Background

This chapter provides essential background to the research. It introduces the declarative mapping lifecycle and describes the main categories of mappings, uplift, ontology alignment, and interlinking. It then discusses metadata and FAIR principles in the context of mapping documentation.

Chapter 3: State of the Art

This chapter presents a comprehensive review of existing work on declarative mappings and metadata management. It examines current approaches to mapping documentation and reuse, along with provenance, quality, and lifecycle ontologies that support these activities. The chapter also explores recent efforts in AI- and LLM-based mapping generation, identifying limitations such as lack of context awareness, inconsistency, and limited reproducibility. These analyses highlight

open challenges and motivate the need for a unified lifecycle-based metadata model.

Chapter 4: Lifecycle-Based Metadata Model

This chapter presents the Lifecycle-Based Metadata Model for Declarative Mappings (LMMD), which constitutes the primary contribution of this thesis. The chapter describes the requirements that motivated the model, the design decisions underpinning its structure, and its formalisation as Mapping Metadata Vocabulary (MMV) ontology that aligned with existing standards. It further demonstrates how the model is operationalised through the MetaSEMAP tool, enabling lifecycle-aware documentation of declarative mappings in practice.

Chapter 5: Evaluation

This chapter reports a series of five studies conducted to evaluate the relevance, completeness, and practical value of the proposed lifecycle-based metadata model. Study 1 presents a community survey assessing metadata requirements. Study 2 evaluates the usability of the model and MetaSEMAP tool using established usability frameworks. Study 3 evaluates the ontology through expert review and automated validation. Study 4 presents a mapping reuse experiment assessing whether lifecycle metadata supports more informed reuse decisions A use case study presented in Chapter 6 introduces the minor contribution of the thesis,, an exploratory experiment on metadata-guided LLM mapping generation, comparing unguided and guided prompts to assess their effect on the semantic correctness and completeness of generated mappings.

Chapter 7: Conclusion

This chapter summarises the overall contributions of the thesis and answers the research question. It reflects on how the research objectives were fulfilled and outlines future research directions, including potential extensions to the metadata

model, a standardised purpose vocabulary for semantic artefacts, and further exploration of metadata-guided mapping generation with LLMs.

Appendices:

The appendices provide supplementary materials that document the design, ontology specification, usability evaluation, and LLM experiments described in this thesis.

Thesis Materials:

All thesis materials including publications, experiments, raw data and results, and source code supporting this thesis are available at the following GitHub repository: https://github.com/SarahAlzahranitcd/Thesis-Materials

Chapter 2: Background

This chapter provides the background necessary for understanding the role of declarative mappings and metadata in knowledge graph and data integration workflows. It introduces the key concepts and technologies underlying the research presented in this thesis. The research presented in this thesis is grounded in three areas: declarative mappings in knowledge graph construction, the concept of the mapping lifecycle, and metadata for mapping documentation and reuse.

The chapter begins by describing the evolution from the Semantic Web vision to the adoption of knowledge graphs for integrating heterogeneous data sources (Section 2.1). Section 2.2 explains the role of declarative mappings in knowledge graph construction pipelines, while Section 2.3 presents the main types of declarative mappings addressed in this thesis.

Section 2.4 introduces the concept of the mapping lifecycle while Section 2.5 discusses the concept of metadata in the context of semantic artefacts and distinguishes between different levels at which metadata can be represented, as well as presenting approaches for representing metadata in RDF. Finally, Section 2.6 introduces the FAIR principles and their relevance to semantic artefacts and mapping documentation.

2.1 From the Semantic Web to Knowledge Graph Integration

The Semantic Web vision, proposed by Tim Berners-Lee and colleagues, introduced a framework to extend the Web from a collection of human-readable resources into a web of data. In this framework information is represented using shared semantic standards [21] that aimed to allow software agents to interpret relationships between entities and integrate information across heterogeneous and distributed resources. Technologies such as the Resource Description Framework (RDF) [22] and the Web Ontology Language (OWL) [23] were introduced to

enable the representation of structured knowledge in machine-readable form and to support data interpretation, processing, and interoperability across systems.

These technologies established the foundations for machine-readable data on the Web. However, practical data integration challenges led to the development of knowledge graphs as a widely adopted approach for organising and integrating heterogeneous data sources. Knowledge graphs model entities and their relationships as interconnected nodes of data, supporting richer capabilities for data integration, querying, and reasoning [24].

Constructing a knowledge graph usually involves integrating data from multiple sources and in different formats or structures such as relational databases, spreadsheets, XML files and web services. These differences in structure and semantics mean that transformation processes are necessary to convert heterogeneous data into a unified semantic representation. Beyond this transformation, ontology alignment may also be required to reconcile different schemas and vocabularies used to describe these resources. Entity interlinking is also needed to connect related entities across datasets represented as separate graphs [1]. In knowledge graph construction workflow, declarative mappings support all three of these activities providing structured and reusable specifications for data transformation, ontology alignment, and entity interlinking .

2.2 Declarative Mappings in Knowledge Graph Construction

Declarative mappings, as discussed in Section 2.1, are the primary method for transforming, aligning, and linking heterogeneous data in knowledge graph construction workflows. Unlike procedural scripts, declarative mappings describe transformation rules as structured specifications that are independent of the execution engine. These specifications make mappings more explicit, reusable, and easier to maintain [3], [5].

These rules define the relationship between elements from the source dataset and concepts from the target dataset. For example, a column in a relational database table may correspond to a class or property in the target semantic model. Transformation engines then automatically generate RDF triples that represent the source dataset within a knowledge graph. A number of standardised languages have been developed to support declarative mapping specifications, such as the W3C standard **R2RML [25]**.

The term *declarative mappings* is commonly associated with data transformation tasks, but other mapping activities exist within knowledge graph ecosystems. In the literature, the term *mapping* is used in different ways depending on the context. Some studies restrict the term to transformation rules that convert non-RDF data into RDF, while others use it more broadly to describe correspondences between ontologies or links between entities across datasets.

In this thesis, the term declarative mappings is used in a broader sense to refer to specifications that declare semantic correspondences between elements of different data or knowledge representations. Under this definition, declarative mappings cover three main types: data transformation mappings (Uplift), ontology alignment mappings, and entity interlinking specifications.

A key observation motivating this broader definition is that semantic correspondences are rarely simple one-to-one class mappings. Scharffe [26] identifies 36 correspondence patterns arranged into a hierarchy of attribute, class, and relation correspondence types, as illustrated in Figure 2.1 [27]. These patterns were originally proposed for ontology alignment. Walshe [28] demonstrates this complexity in Linked Open Data, showing that correspondences between datasets such as DBpedia and YAGO2 require complex pattern-based representations rather than simple class-to-class mappings. The variety of patterns identified reflects the same semantic relationships that arise in uplift mappings and interlinking, suggesting that correspondence patterns have the potential to capture correspondences relevant across all three mapping types addressed in this thesis.

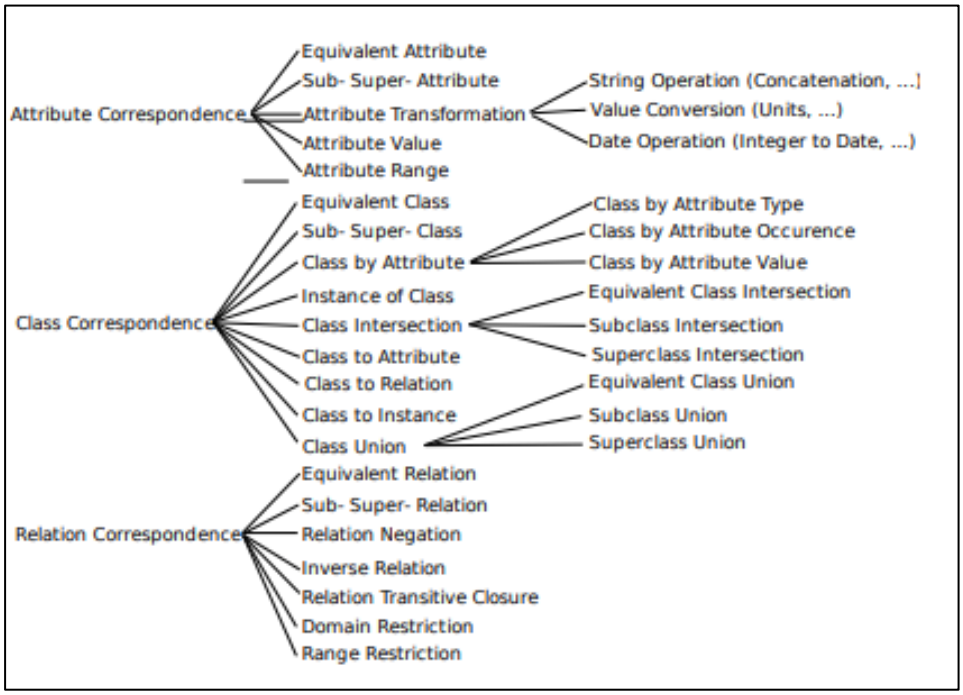


Figure 2.1 Organisation of basic correspondence patterns [27]

Beyond the complexity of correspondences, declarative mapping types share further common characteristics. All three result in explicit, machine-readable specifications that are independent of the execution engine, although their syntactic representations differ. These declarative mappings are executed by processors or engines to support data processing tasks such as transforming, aligning, or linking datasets or schemas. The processes involved in creating these mappings also share similar phases, which are discussed in detail in Section 2.4.

2.3 Types of Declarative Mappings

Within Linked Data and Semantic Web domains, mapping activities can be classified into three main types: **data transformation**, **ontology alignment**, and **interlinking**. The following subsections introduce the main types of mappings considered in this thesis.

2.3.1 Data transformation mappings

Data transformation mappings describe how data from structured or semi-structured sources are transformed into RDF representations or vice versa. Two terms describing data transformation processes can be found in the literature: uplift and downlift mappings [3]. The uplift mapping process involves converting data from different formats into RDF, while the downlift mapping process refers to the reverse transformation from RDF to other formats. The type of mapping considered in this thesis is uplift mapping, which is commonly used in knowledge graph construction and data integration workflows. Downlift mappings are outside the scope of this thesis.

Several declarative mapping languages have been developed to support the uplift process. One of the most widely used languages for expressing uplift mapping is the W3C standard R2RML which provides a mechanism for transforming relational databases into RDF [25]. Several extensions to R2RML have been proposed, such as RML [3], which extends R2RML to support a wider range of input formats including CSV, JSON, and XML, and is currently undergoing standardisation through W3C Knowledge Graph Construction Community Group [29], [30]. Mappings expressed using R2RML and its extensions are represented as RDF documents, which allows them to be shared, queried and annotated.

The structure of these mapping documents consists of one or more Triples Maps, which define how RDF triples are generated from an input source. A Triples Map typically includes a Logical Source, a Subject Map, and one or more Predicate-Object Maps. These components define how elements from the source dataset are mapped to RDF triples in the target knowledge graph. Figure 2.2 shows a simplified example of an RML mapping specification illustrating the structure of a Triples Map.

```
@prefix rr: <http://www.w3.org/ns/r2rml#>.
@prefix rml: <http://semweb.mmlab.be/ns/rml#>.
@prefix ex: <http://example.com/ns#>.
@prefix ql: <http://semweb.mmlab.be/ns/ql#>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@base <http://example.com/ns#>.

<#TransportMapping> a rr:TriplesMap;
  rml:logicalSource [
    rml:source "Transport.xml" ;
    rml:iterator "/transport/bus/route/stop";
    rml:referenceFormulation ql:XPath;
  ];

  rr:subjectMap [
    rr:template
      "http://trans.example.com/stop/{@id}";
    rr:class ex:Stop
  ];

  rr:predicateObjectMap [
    rr:predicate rdfs:label;
    rr:objectMap [
      rml:reference "."
    ]
  ].
```

Figure 2.2 Example of an RML mapping specification illustrating the structure of a Triples Map [30].

The primary components of a Triples Map are as follows:

- The Logical Source specifies the input dataset and how it is accessed.
- The Subject Map specifies how the subject of the RDF triple is formed using a URI template. It also specifies the type of the generated resource.
- A Predicate-Object Map defines how predicates and objects are generated. Predicate Map specifies the predicate of the triple whereas the Object Map defines how object values are produced from the source data.

While data transformation mappings focus on converting heterogeneous data into RDF representations, other mapping activities address semantic interoperability across knowledge graphs such as ontology alignment and interlinking.

2.3.2 Ontology Alignment mappings

In knowledge graph integration scenarios, datasets are often described using different ontologies. Thus, semantically similar concepts may be represented using different classes or properties. To address this, ontology alignment has been proposed as a method of identifying correspondences between concepts in different

ontologies [2]. These correspondences facilitate data integration and enable semantic interoperability across knowledge graphs.

Ontology alignment specifications are often represented using structured formats. This structured format describes the entities involved in the alignment process and specifies the type of relationship between them. Relationships are usually expressed as equivalence, subsumption, or similarity between classes or properties. The results of ontology alignment are typically stored as alignment files, which can be processed by alignment tools or used to support integration tasks across datasets. Similar to the uplift mapping, these files are typically represented as RDF documents.

A widely recognised initiative supporting the evaluation of ontology alignment systems is the Ontology Alignment Evaluation Initiative (OAEI) [16]. The OAEI provides benchmark datasets and evaluation campaigns that allow researchers to compare the performance of ontology matching systems across different domains. Each year, participating systems are evaluated on a set of predefined tasks using metrics such as precision, recall, and F-measure.

Figure 2.3 shows a simplified example of an alignment specification expressed using INRIA's Alignment API format illustrating a correspondence between two ontology classes. The example is derived from the output of the LogMap ontology alignment system, one of the systems evaluated in the OAEI Anatomy track. In this example, a class from the Mouse Anatomy ontology is aligned with a corresponding class in the Human Anatomy ontology with an associated confidence score.

```xml
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF xmlns="http://knowledgeweb.semanticweb.org/heterogeneity/alignment"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#">

<Alignment>
<xml>yes</xml>
<level>0</level>
<type>??</type>
<onto1>http://mouse.owl</onto1>
<onto2>http://human.owl</onto2>
<uri1>http://mouse.owl</uri1>
<uri2>http://human.owl</uri2>
<map>
    <Cell>
        <entity1 rdf:resource="http://mouse.owl#MA_0002668"/>
        <entity2 rdf:resource="http://human.owl#NCI_C32827"/>
        <measure rdf:datatype="xsd:float">0.7</measure>
        <relation>=</relation>
    </Cell>
</map>
<map>
    <Cell>
        <entity1 rdf:resource="http://mouse.owl#MA_0000378"/>
        <entity2 rdf:resource="http://human.owl#NCI_C12416"/>
        <measure rdf:datatype="xsd:float">0.94</measure>
        <relation>=</relation>
    </Cell>
</map>
```

Figure 2.3 Example of an ontology alignment correspondence expressed using the Alignment API format [31].

2.3.3 Entity Interlinking mappings

In knowledge graph integration scenarios, the same real-world entity may appear in various datasets using distinct identifiers or naming conventions. For example, the city of Paris may be represented as *dbpedia:Paris* in DBpedia and as *geonames:2988507* in the GeoNames dataset. Even though both identifiers refer to the same real-world entity “Paris” , they originate from different datasets and use different URIs. Without explicit links between these identifiers, integrating or querying information about the same entity across multiple knowledge graphs becomes difficult.Entity interlinking addresses this by identifying similar or related entities representing the same real-world object across different datasets [1].

Entity interlinking are typically expressed using link specifications that define the conditions under which two entities should be considered equivalent or related. These specifications use similarity measures, comparison functions, or transformation rules to link entity attributes. Tools such as Silk and LIMES allow users to define these link specifications declaratively in order to automatically

generate links between datasets [31], [32]. In some cases, simple interlinking conditions can also be expressed directly as SPARQL queries [33] that identify matching entities across datasets based on shared attributes. Similar to the previously described declarative types, the result of the declarative interlinking process is represented as RDF triples that express relationships between entities.

A common example of interlinking is the use of the *owl:sameAs* property to show that two URIs refer to the same entity. Figure 2.4 presents a simplified version of a Silk link specification used to generate *owl:sameAs* links between entities from DBpedia and GeoNames. The specification outlines the source and target datasets, together with a similarity condition based on the comparison of entity labels using the Jaro string similarity metric.

```
<Interlink id="cities">
  <LinkType>owl:sameAs</LinkType>

  <SourceDataset dataSource="dbpedia" var="a">
    <RestrictTo>{ ?a rdf:type dbpedia:City }</RestrictTo>
  </SourceDataset>

  <TargetDataset dataSource="geonames" var="b">
    <RestrictTo>?b gn:featureClass gn:P</RestrictTo>
  </TargetDataset>

  <LinkCondition>
    <Compare metric="jaroSimilarity">
      <Param name="str1" path="?a/rdfs:label"/>
      <Param name="str2" path="?b/gn:name"/>
    </Compare>
  </LinkCondition>
</Interlink>
```

Figure 2.4 Example of a Silk link specification for entity interlinking (Volz et al., 2009).

Despite their differences in purpose and representation, the three mapping types share notable structural characteristics. Table 2.1 illustrates these parallels across uplift mappings, ontology alignment, and entity interlinking. As shown in Table 2.1, while the three mapping types differ in their source and target elements, correspondence types, and execution engines, they share a common underlying structure. All three express semantic correspondences between elements of

different data or knowledge representations as machine-readable specifications executed by dedicated processing engines [1], [2], [3].

Specification formats used across mapping types differ considerably. Uplift mappings have converged around the W3C standards R2RML [25] and RML [3], which is under a standardisation process, with YARRRML [15] providing a more human-readable syntax that compiles to RML. Ontology alignment has historically relied on the Alignment API format [34] and its expressive extension EDOAL [35] for complex correspondences, while SSSOM [14] has emerged as a simpler, FAIR-oriented alternative. Entity interlinking specifications are primarily expressed using tool-specific languages such as Silk LSL [31] and LIMES configuration files [32], or directly as SPARQL CONSTRUCT queries, reflecting a less standardised landscape compared to the other two mapping types.

Despite this diversity in representation, all three mapping types follow the same lifecycle pattern from analysis and design through to development, testing, and maintenance [5]. This observation directly motivates the unified lifecycle-based metadata model proposed in this thesis.

Table 2.1 Comparative overview of declarative mapping types across key characteristics

Characteristics	Uplift Mapping	Ontology alignment.	Interlinking
Source element	Data field (CSV column, DB table, JSON attribute)	Ontology class or property	Entity URI
Target element	Ontology term or property	Ontology class or property	Entity URI
Correspondence type	Transformation rule	Equivalence, subsumption, or similarity	owl:sameAs, skos:closeMatch
Specification format	R2RML , RML , YARRRML	Alignment API , EDOAL, SSSOM, OWL axioms, SKOS	Silk LSL, LIMES config, SPARQL CONSTRUCT queries
Output	RDF triples	Alignment file (RDF/XML or TSV)	RDF triples expressing entity links
Execution engine	R2RML and RML processors, Morph-KGC	LogMap, AML, FCA-Map	Silk, LIMES

The specific forms of semantic correspondence within each mapping type can be further characterised using correspondence patterns [26], [27]. Table 2.2 illustrates how some example correspondence patterns identified by Scharffe and Fensel [26] might appear across all three declarative mapping types.

Table 2.2 Examples of correspondence patterns across the three declarative mapping types

Correspondence pattern	Uplift Mapping	Ontology alignment.	Interlinking
Attribute Value Transformation	CSV column date_of_birth mapped to foaf:birthday after converting date format to ISO 8601	Property dateOfBirth in ontology A aligned to birthDate in ontology B after applying a date normalisation transformation	Entity labels normalised before string similarity comparison in Silk LSL
Class by Attribute Value	Rows where type = "Person" mapped to foaf:Person class	Class Human in ontology A equivalent to class Person in ontology B when attribute species = "homo sapiens"	Entities with rdf:type = "Person" linked across DBpedia and Wikidata
Attribute Similarity	Free-text field values matched to controlled vocabulary terms using string similarity	Confidence score computed using string or semantic similarity between class labels	Jaro similarity applied to entity labels to generate owl:sameAs links between DBpedia and GeoNames

As shown in Table 2.2, correspondence patterns originally proposed for ontology alignment [26] also appear in uplift mappings and entity interlinking, expressed through different languages and tools. This observation suggests that all three mapping types express variants of the same underlying semantic correspondences, a characteristic that is revisited in Chapter 3 when examining the limitations of existing metadata approaches.

2.4 The Mapping Lifecycle

Declarative mappings are often shared as standalone artefacts. Their creation typically results from multiple stages. Similar to other data integration and software development workflows, mapping development involves several activities that take place before, during, and after the implementation of the

mapping specification [5]. The term mapping lifecycle is used to describe the sequence of stages involved in developing these mappings [5], [36].

Figure 2.5 illustrates a general software development lifecycle [37], which serves as the basis for the mapping lifecycle adopted in this thesis.

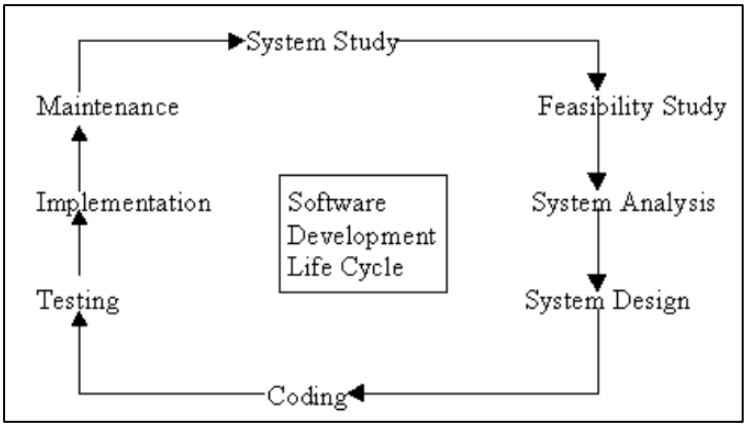


Figure 2.5 General software development lifecycle phases[38]

In practice, the development of mappings usually begins with an analysis stage (**Analysis**). This stage involves specifying stakeholders involved and identifying the scope and purpose of the mapping project, including its requirements, assumptions, risks, and the source and target datasets or ontologies. These elements are examined at the beginning of the mapping project in order to understand its scope, the structure of the data sources, and their semantic compatibility. Based on this understanding, the mapping design stage takes place (**Design**), during which the relevant correspondences are defined and decisions about whether to create new mapping from scratch or reuse of existing mappings may also be considered.

Following the design stage, the mapping specification can be implemented (**Development**) using declarative mapping languages or linking rules. After that, the generated results must be evaluated (**Testing**). This evaluation involves verifying the generated triples, checking for syntactic or semantic errors, assessing

the overall quality of the mapping, and checking that requirements of the project are met.

Mappings also require maintenance (**Maintenance**) as final and ongoing activity in the lifecycle. Typically the underlying datasets of the mapping, schemas, or ontologies evolve over time. Changes in data structures, ontology versions, or integration requirements may require updating or extending existing mappings. Mappings can quickly become outdated and produce incorrect or incomplete results without proper and regular maintenance. Maintaining mappings ensures the long-term usability of mappings and the knowledge graphs that depend on them.

Despite the importance of these lifecycle activities, existing mapping artefacts typically capture only the final mapping specification, with limited information describing how the mapping was created. Important contextual information about why a mapping was developed and how it was designed, evaluated, and maintained over time is often missing or insufficiently documented. Therefore, existing mappings shared on the web may lack the information needed to understand their purpose, assess their quality, or determine whether they can be reused in another context.

Several researchers have proposed mapping lifecycle models in response to these challenges [5], [38], [39], [40]. These models aim to organise mapping activities into well-defined stages and provide guidance on how mappings should be created, evaluated, and maintained throughout their lifecycle. Chapter 3 reviews existing approaches to mapping lifecycles and discusses how lifecycle thinking has evolved to support better organisation of the mapping activities, documentation and their reuse.

2.5 Metadata in the Context of Declarative Mappings

Metadata plays an important role in supporting documentation, discovery, and reuse of declarative mappings. Mapping specifications usually focus on describing how data elements across datasets or ontologies correspond to each other, while

metadata focuses on providing more context and explanation about the mapping, for example how they were created and for what purpose [4].

Mappings are often shared with little or no accompanying metadata, which makes it difficult to understand or reuse them. Metadata can address this challenge by describing the characteristics of mapping artefacts such as their purpose, provenance, development details, and evaluation.

The following subsections introduce the concept of metadata for semantic artefacts, discuss different levels at which mapping metadata can be captured, and present the RDF approaches commonly used to represent metadata.

2.5.1 Defining Metadata for Semantic Artefacts

Metadata is often referred to as data about data [10]. In the context of semantic web technologies, metadata is used to describe semantic artefacts such as datasets, ontologies, and mappings. This can include information about provenance, usage, development, and stakeholders involved. For metadata to support automated workflows, it must be represented in a structured, machine-readable format rather than captured informally.

For declarative mappings in particular, Shvaiko and Euzenat [2] emphasise the need to carefully consider alignment metadata and annotations in order to provide standard ways to communicate and retrieve alignment mappings.
Other research efforts have proposed metadata models to enrich mappings with descriptive information to improve transparency, interoperability, and FAIR alignment practices, which are reviewed in detail in Chapter 3.

2.5.2 Levels of Mapping Metadata

Metadata describing mappings can be captured at different levels of granularity. In many existing systems, metadata is defined after the mapping is developed or published; therefore, it is often attached at the dataset or graph level, which can also be seen as a project-level description of the semantic artefact [4].

However, metadata can also be captured at finer levels of granularity. For example, metadata can be represented at the triple level, where each RDF triple is described individually. This can be seen clearly in ontology alignment, where each correspondence is annotated with measures such as confidence scores. Metadata can also be captured at the term level, where each concept or term in an RDF file is described or annotated individually.

In this thesis, the goal is to provide a metadata model that is applicable across different semantic mapping activities, independent of their implementation details and representation. The project level is therefore adopted as the primary level of metadata capture in this thesis, as discussed in Chapter 4.

2.5.3 Metadata Representations in RDF

Representing metadata in RDF involves writing RDF statements about existing RDF statements. Several approaches have been proposed for this purpose including standard reification, named graphs, singleton property, and RDF-star, each with different trade-offs in terms of expressiveness, scalability, and query performance [41].

RDF reification involves describing an RDF triple using four additional statements. This description uses vocabulary such as *rdf:Statement*, *rdf:predicate*, and *rdf:object*. However, this approach has two main drawbacks: it lacks formal semantics and requires four additional triples per one statement, which significantly increases dataset size [41].

The Singleton Property approach [42] was proposed to address some of these limitations by creating a unique predicate for each statement, allowing metadata to be attached through additional triples. This reduces the number of triples required and preserves the original structure of the data, but generates a large number of unique predicates, which can be difficult to manage [42].

Another approach is RDF-star [43], [44], which extends RDF syntax to allow nesting within the graph by embedding triples using << and >>. This representation

results in smaller documents, faster data exchange, and simpler SPARQL queries using SPARQL-star, making it easier to understand and use compared to standard reification [44].

Named Graphs [45] provide another way to represent metadata. In this approach a group of RDF triples is assigned a unique identifier. This allows metadata to be attached to the entire graph rather than individual triples.

Several studies have evaluated these approaches based on criteria such as number of triples, storage size, and query execution time [46],[41]. Results show that Singleton Property reduces the number of triples compared to standard reification, while RDF-star achieves even greater reduction. In terms of execution time, RDF-star performs the best, while Singleton Property performs the worst. Additional benchmarking using Wikidata datasets also shows that RDF-star provides better performance in terms of number of statements, loading time, and repository size, while Named Graphs produce fewer statements compared to other approaches.

2.6 FAIR Principles and Semantic Artefacts

Metadata provides a structured description of mapping artefacts. It can describe mapping characteristics, provenance, domain and purpose in a structured and machine-readable way. This description is important for managing mappings and enabling their discovery and reuse. Despite their use in a wide range of applications, researchers still face challenges in locating, accessing, understanding, and reusing mappings. These challenges are mainly caused by the lack of detailed documentation and metadata [9].

To address these limitations, the FAIR (Findable, Accessible, Interoperable, Reusable) principles provide a framework for improving the management and reuse of digital artefacts [47]. Applying FAIR principles to declarative mappings requires structured metadata and proper documentation throughout their lifecycle. Recent initiatives have also highlighted the need to treat mappings as first-class

artefacts and ensure their FAIRness, which is discussed in more detail in Chapter 3.

The Findable principle ensures that mappings and their metadata can be discovered by both humans and machines by the use of unique identifiers. The Accessible principle requires clear conditions for accessing mappings once they are found. The Interoperable principle enables integration across systems by using shared and standard representations. Finally, the Reusable principle emphasises the need for sufficient metadata describing the mapping purpose, provenance, and usage conditions.

Achieving FAIR principles for declarative mappings depends on providing metadata that describes not only the final mapping, but also the processes involved in its development and maintenance. This requirement directly motivates the lifecycle-based metadata model proposed in this thesis.

2.7 Chapter Summary

This chapter introduced the key concepts needed to understand the research presented in this thesis. It began by tracing the evolution of the web from the Semantic Web to knowledge graph construction and integration. The role of declarative mappings in transforming, aligning, and interlinking heterogeneous data was covered, along with an overview of the main mapping types. The mapping lifecycle was then introduced with its five phases. This was followed by the discussion of metadata, its definition, levels of granularity, and RDF representations. Finally, the chapter discussed the FAIR principles and their relevance to declarative mappings. All these concepts establish the foundation for the existing approaches in Chapter 3.

Chapter 3: State Of The Art

This chapter examines related work identified through the state of the art review in relation to the research questions and objectives of this thesis. The literature can be organised around three interrelated aspects. The first aspect (Section 3.1) reviews mapping lifecycle frameworks and analyses how they have evolved since their initial proposals. The second aspect (Section 3.2) examines current practices for documenting and presenting declarative mapping metadata, with particular attention to metadata granularity and lifecycle coverage. Section 3.3 discusses emerging AI-based mapping generation approaches and analyses how the absence of contextual metadata may affect their reliability. The organisation of this chapter follows this progression, moving from lifecycle frameworks to metadata documentation, and then to AI-based generation. Together they identify the limitations and open challenges that motivate the research in this thesis.

The reviewed work in this chapter was identified through a structured search using Google Scholar, which indexes sources such as the ACM Digital Library, IEEE Xplore, and Semantic Scholar. The search began with a key relevant paper and expanded through backward and forward citation tracking. Retrieved studies were filtered based on their relevance to the scope of this thesis.

3.1 Approaches to Structuring the Mapping Lifecycle

Several lifecycle models have been proposed to structure the development and management of declarative mappings. These models aim to organise mapping activities into defined phases and provide guidance for mapping creation, validation, and maintenance. As mapping artefacts have grown in complexity and importance within data integration pipelines, the need for structured lifecycle governance has become increasingly evident. However, the scope, terminology, and level of detail differ across proposals. This section reviews these models, tracing their initial introduction and subsequent evolution over time, in order to

understand how lifecycle thinking has developed within declarative mapping engineering.

3.1.1 Early Lifecycle Models

Early efforts to structure the mapping lifecycle emerged from the community's need for a commonly agreed-upon lifecycle model for ontology mapping at a time when the field was still relatively young compared to ontology engineering and database schema mapping [38]. As alignment tasks became more complex and widely used, particularly within evaluation initiatives, researchers began to propose explicit lifecycle models, emphasising that mapping is a multi-phase process rather than a set of isolated technical steps.

One of the first foundational contributions to lifecycle modelling in ontology mapping was proposed in the **OISIN framework** [38], which introduced an explicit four-phase ontology mapping lifecycle. The proposed phases include the characterisation phase, the mapping phase, the execution phase, and the management phase.

In this model, **the characterisation phase** identifies the ontologies involved, analyses them, and recognises potential difficulties in the matching process. It also supports the generation of candidate matches. **The mapping phase** then covers the activities involved in validating and formalising the mappings based on the candidate matches identified previously. **The execution phase** concerns the rendering of the confirmed mappings in a suitable format. Finally, **the management phase** is responsible for the ongoing maintenance and evolution of the mapping after deployment.

Figure 3.1 shows how these phases are connected and represented. The diagram illustrates how ontology discovery and characterisation activities take place before deciding to match, followed by the selection and application of matching algorithms and the identification of confirmed mappings. It also shows decision points within the lifecycle, indicating that the process is not strictly linear and that

returning to a previous phase is possible. The diagram also includes a datastore for mapping information, showing how rendering, sharing, altering, and integrating mappings are connected processes within the execution and management phases. This representation highlights that ontology mapping is a coordinated and iterative process that must be managed over time.

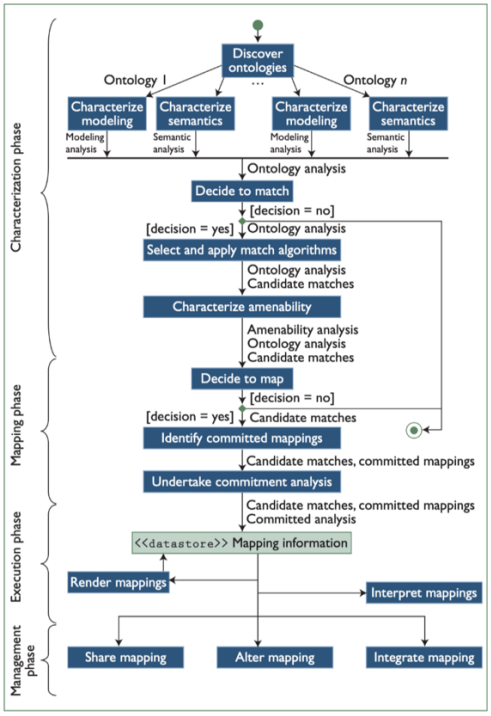


Figure 3.1 OISIN ontology mapping lifecycle represented as a UML activity diagram [38]

While the OISIN framework introduced a structured four-phase lifecycle for ontology mapping, later work by Thomas et al. [39] expanded lifecycle granularity and placed a stronger emphasis on metadata documentation. In their proposed **OM2R model**, a clearer separation between matching and mapping activities was introduced, distinguishing the generation of correspondence candidates from the validation and confirmation of the actual mappings. This separation and the additional phases introduced in this proposal allowed a more precise representation of the mapping lifecycle activities.

The main design objective of the OM2R framework was to create a metadata model for ontology mappings that covers the complete lifecycle of the mapping. The authors argued that the key understanding of how a mapping is created lies in the ontology mapping lifecycle; therefore, the individual phases of the lifecycle were used as the structure for their proposed OM2R metadata model (discussed in Section 3.2). This lifecycle consists of six phases: the characterisation phase, the matching phase, the mapping phase, the execution phase, the management phase, and the metadata creation phase.

The characterisation phase in this work is described as the phase that includes activities related to ontology discovery. This lifecycle also differs from OISIN by distinguishing between the matching and mapping phases, where the matching phase is introduced as a more granular phase preceding the mapping phase. The main objective of **the matching phase** is the identification of mapping candidates, either manually or automatically, while **the mapping phase** focuses on the creation of confirmed mappings. **The execution phase** in this lifecycle is similar to that introduced in OISIN, involving the rendering of confirmed mappings into different formats to enable processing and sharing. **The management phase** is also similar to that in OISIN, focusing on managing and maintaining the produced mappings. A newly proposed phase in this lifecycle, besides the matching phase, is **the metadata creation phase**, which the authors describe as conceptually a parallel activity to all other phases in the framework, where metadata are collected and processed. The following diagram (Figure 3.2) summarises the lifecycle proposed in this work.

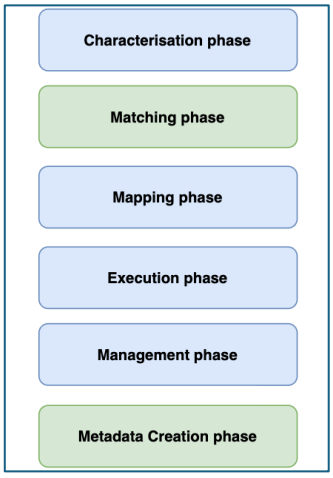


Figure 3.2 OM2R ontology mapping lifecycle illustrating six phases and parallel metadata creation

Although early lifecycle models such as OISIN and OM2R introduced important structuring of ontology mapping activities, they were primarily limited to the ontology alignment community. Their main focus was on formalising lifecycle phases and strengthening documentation practices within that research context. While the OM2R model introduced important refinements, including a clearer separation between matching and mapping and an explicit emphasis on metadata creation, its scope was still primarily limited to ontology alignment scenarios. Neither OISIN nor OM2R addressed mapping types beyond ontology alignment, uplift mappings and entity interlinking were not considered within their lifecycle frameworks.

Broader governance considerations, such as persistent identification of mappings, repository integration, and large scale reuse across communities, were not explicitly addressed in those models. These aspects became more relevant as mapping activities expanded beyond controlled evaluation settings and into collaborative and institutional environments. Although these early models provided a strong conceptual foundation for lifecycle thinking, they did not develop into broadly adopted approaches across different mapping types, domains, and infrastructures. This shift in needs gradually led to governance-oriented

perspectives that framed mapping not only as a structured technical process but also as a managed project-level activity.

3.1.2 Project-Centric Governance Models

As ontology mapping activities became part of broader integration and linked data projects, the need for governance oriented lifecycle approaches became more evident [5]. Early lifecycle models focused primarily on structuring the technical phases of the ontology alignment process, with attention to the mapping artefact itself. In contrast, project-centric governance models frame the ontology mapping as a managed project involving stakeholders, requirements, decisions, and long-term maintenance.

Debruyne et al. in [5] argue that existing metadata approaches tend to focus on alignments and correspondences, leaving out the broader project context. They propose a project-centric view that captures the full lifecycle of a mapping project, including stakeholder activities, scope definition, requirement formulation, and evaluation processes. This project-centric view enables more governance queries, and support mapping discovery, reuse, and traceability.

To support traceability and transparency, the authors proposed modelling the mapping creation as a project from its inception to execution, adopting provenance-based concepts to record activities. The proposed framework builds upon PROV-O [11] to represent the interaction between stakeholders and the mapping phases that lead to the mapping artefact. The proposed lifecycle in the work is based on earlier lifecycle models but takes a project-centric view into account. This lifecycle consists of six phases: stage, characterise, reuse, match, align and map, and application.

During **the stage phase**, stakeholders involved in the mapping project are identified, and the scope of the project and requirements are defined. This is a newly introduced phase compared to previous lifecycles and reflects the project-

centric perspective. **The characterise phase** is concerned with the discovery and analysis of the ontologies to be mapped and the identification of potential difficulties. **The reuse phase** then takes place, which is also newly introduced, as it integrates reuse activities as part of the mapping workflow. During this phase, existing alignments are assessed to determine whether they are suitable for immediate reuse, could serve as a basis for a new alignment, or whether no suitable alignment exists and a new one should be created from scratch.

The match phase then follows, which is quite similar to the matching phase in OM2R lifecycle models. This phase concerns the planning, execution, and evaluation of ontology matcher applications, involving activities such as agreeing on matcher configurations, generating candidate correspondences, and evaluating them for fitness. T**he align and map phase** then takes place, which is similar to the mapping phase in earlier models and is concerned with the creation of the alignment and the mapping. The mapping is defined as the artefact that software agents are able to execute. Finally, **the application phase** focuses on the use of mappings rendered in the previous phases to support the application foreseen by the community stakeholders. This can be considered a newly added phase, supporting the idea of providing the community with new insights about the mapping, for example when the structure of data sources changes, which may lead to a new version of the mapping (change detection).

As illustrated in Figure 3.3, the lifecycle explicitly incorporates stakeholder involvement, requirement formulation, and reuse assessment as integral parts of the mapping workflow. This reflects the shift from viewing mappings purely as technical artefacts toward understanding them as outcomes of structured project activities.

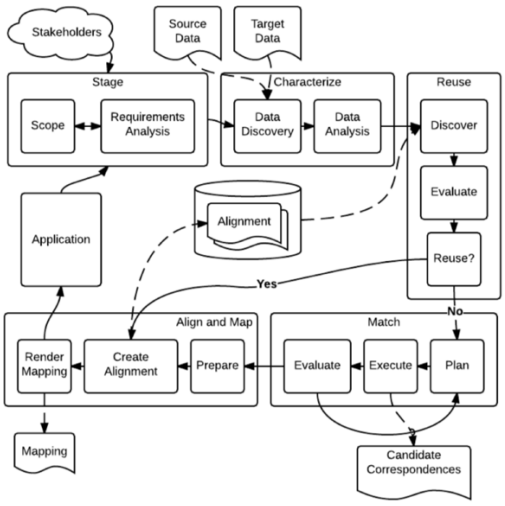


Figure 3.3 Project-centric ontology mapping lifecycle proposed by Debruyne et al.[5]

While project-centric governance models extended lifecycle thinking beyond the mapping artefact, their scope remained primarily focused on ontology mapping scenarios. Similar to OISIN and OM2R, the applicability of this model to uplift mappings and entity interlinking specifications was not explicitly addressed.

Challenges related to findability, interoperability, and reuse remained unresolved at a broader scale. As mappings increasingly functioned as key components for semantic interoperability within large data ecosystems such as the European Open Science Cloud (EOSC), limitations in discoverability and reuse became more evident [9]. Existing mappings were often difficult to locate, embedded within artefacts, or represented in non-interoperable formats, resulting in repeated implementation efforts [9]. This analysis highlighted that mappings themselves were not fully aligned with FAIR principles.

3.1.3 FAIR-Oriented Governance Initiatives

Mappings have increasingly become recognised as key components of semantic interoperability within large-scale data infrastructures, extending their role beyond individual projects. In such contexts, mappings are no longer viewed as project outputs but as reusable semantic artefacts that must be discoverable and

interoperable across domains. This shift in perspective has led to FAIR-oriented initiatives that emphasise treating mappings as digital objects that should comply with the FAIR principles.

FAIR-IMPACT

One of the most significant of these initiatives is FAIR-IMPACT, a Horizon Europe project that commenced in June 2022 [48], with the aim of expanding FAIR solutions across the European Open Science Cloud (EOSC). Within this project, mappings are recognised as first-class artefacts. A dedicated task focused specifically on establishing guidelines and methodology for creating, documenting, and sharing FAIR mappings [9], [40]. The initiative engaged with a wide range of communities through workshops, survey, and events to gather mapping practices and requirements across domains.

As part of their recommendations on how to make mappings FAIR, they proposed a Mapping Practice Framework. This framework is defined as a proper method for creating and curating mappings from a general perspective. It is intended to be applicable across different types of mappings, unlike previous lifecycle models that were mainly tailored to a specific mapping type.

The proposed framework consists of five phases: pre-mapping, mapping, review, hosting, and maintenance. **The pre-mapping** phase is described as the foundational phase and is similar to the stage phase in the previous project-centric lifecycle; during this phase, discussions with stakeholders take place, and the source and target artefacts for mapping are defined. **The mapping phase** combines both match and map phases from previous lifecycles and represents the practical implementation of the mapping. **The review phase** in this work is described as the assessment of the applicability and quality of the mapping exercise; depending on the reviewer's decision, a mapping file can either be accepted and moved to the next phase or require further refinement. **The hosting phase** involves uploading the mapping file to a hosting resource along with its associated metadata. Finally,

the maintenance phase targets the long-term availability and accuracy of the mapping artefact.

As discussed in their work, the proposed framework has been tested by different communities with further testing ongoing. Figure 3.4 illustrates the five phases of the Mapping Practice Framework and the distribution of stakeholder roles across them. The diagram emphasises the structured progression from pre-mapping to maintenance, as well as the collaborative and iterative nature of the mapping process.

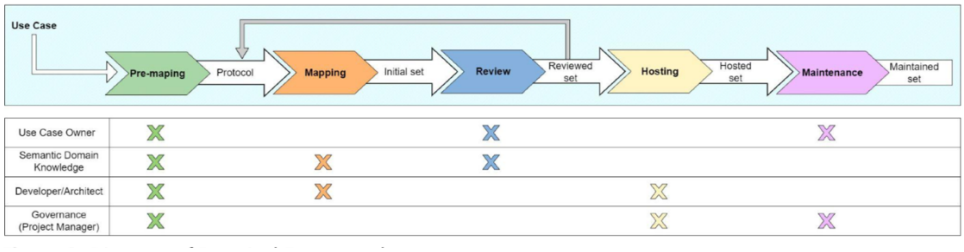


Figure 3.4 FAIR-IMPACT Mapping Practice Framework illustrating five phases and stakeholder involvement[9], [40]

RDA FAIR Mappings WG

Building on and extending FAIR-IMPACT work beyond Europe, the Research Data Alliance (RDA) FAIR Mappings Working Group was endorsed in February 2025 and runs until August 2026 [18]. The Group aims to produce common guidelines for making different types of mappings FAIR. Its expected outputs include technical recommendations for FAIR by design mappings, a methodology covering the mapping lifecycle, a unifying classification of mapping types, and a knowledge base of mapping scenarios [18].

The group has published initial outputs including collected case studies from diverse communities, a conceptual model, and mappings schema [18]. However, its full technical recommendations and unified metadata model remain under development.

While these FAIR-oriented initiatives recognise mappings as reusable semantic artefacts that must be discoverable and maintained within broader data ecosystems, neither explicitly defines how lifecycle phases and metadata requirements differ across different types of mappings. Neither provides a concrete, machine-readable metadata model applicable across all mapping types. The question of how declarative mapping metadata should be structured, represented, and exchanged in practice across different mapping types and lifecycle stages remains open, a gap that Section 3.2 examines in detail.

3.2 Approaches to Documenting Mapping Metadata

Beyond lifecycle and governance frameworks, effective reuse and interoperability also depend on what metadata is captured and how mapping metadata is documented and represented. Different approaches have been proposed to capture mapping information. This section reviews these approaches and analyses their scope, granularity, and limitations.

3.2.1 Lifecycle-Based metadata models

Two lifecycle-based metadata models have been proposed in the literature that organise metadata according to mapping lifecycle phases: the OM2R metadata model [39], and the project-centric governance metadata model proposed by Debruyne et al [5].

OM2R Metadata Model

An early example of a lifecycle-based metadata model is OM2R, which organises metadata according to the phases of the ontology mapping lifecycle. The model lifecycle is reviewed in Section 3.1.1 [39].

OM2R built its metadata model based on the proposed lifecycle and provides a formal vocabulary for the full ontology mapping lifecycle, covering

characterisation, matching, mapping, execution, and management. The model is formally expressed in OWL-DL, consisting of 38 classes and 21 typed object relations. In total, the model introduced 29 metadata fields categorised across the lifecycle to provide a common vocabulary for machine-based discovery and retrieval [40].

In **the identification and characterisation phases**, the model captures metadata describing the ontologies involved in the mapping. These include ontology name and description, location (URI/URL), unique identifier, version, and creation date, as well as editor or contact information. Structural and technical characteristics are also documented, such as the formal language (e.g., OWL-DL), exchange notation (e.g., RDF/XML), ontology size indicators (number of classes, properties, and instances), and basic design classifications (e.g., flat or deep hierarchies).

For **the matching phase**, OM2R provides more detailed documentation of how correspondences were generated. The model includes metadata fields for the location of the alignment file, the matching language and notation, and the matching method (manual, automatic, or mixed). It further allows documentation of the specific matching tool and algorithm used, including implementation details such as Java class names or source code references. Additional parameters such as applied similarity thresholds, matching scope (complete or partial), analysed elements (e.g., labels), and specific matching requirements are also represented.

At **the mapping and execution levels**, metadata can describe confirmed mappings, rendering formats, and storage references, while the management phase includes information relevant to sharing, versioning, and maintenance of mappings.

OM2R provides relatively fine-grained lifecycle metadata for ontology alignment scenarios. However, its granularity and terminology are often closely aligned with ontology alignment workflows. As a result, its applicability to other mapping types, such as data transformation mappings or interlinking specifications, is not always explicit.

Table 3.1 summarises the key metadata fields captured by OM2R across its lifecycle phases. The full field specifications including implemented ontology property names are provided in Appendix A, Section A.1.

Table 3.1 Key metadata fields captured by OM2R

Phase	Key Metadata Fields
Characterisation	Ontology name, description, location, identifier, version, creation date, editor, formal language, notation, size indicators, design classification
Matching	Alignment location, matching language, notation, method, tool, algorithm, threshold, scope, analysed elements, requirements
Mapping/Execution	Confirmed mappings, rendering formats, storage references
Management	Version identifier, hash ID, change log
Metadata Creation	Parallel activity across all phases

The Project-Centric Governance Metadata Model

Project-centric governance models also adopt a lifecycle-based approach to metadata documentation. However, unlike OM2R, which focuses primarily on documenting alignment lifecycle activities, the project-centric model extends metadata coverage to include governance and project-level information [5]. As discussed in Section 3.1.2, in this approach a mapping is not treated as a technical artefact, but as the outcome of a managed project involving stakeholders, requirements, and evaluation processes.

In the **stage phase**, metadata captures the inception of the mapping project, including the involved community of stakeholders, individual stakeholder roles, the defined scope of the project, and formally represented requirements that the final alignment must satisfy.

During the **characterisation phase**, metadata documents ontology discovery and assessment activities. This includes discovery reports, ontology analysis summaries describing quality and suitability, selection of the specific source and target ontologies, and a formal mapping decision confirming whether to proceed.

The **reuse phase** explicitly integrates evaluation of pre-existing alignments. Metadata artefacts include reuse evaluation reports, reuse decisions (direct reuse, adaptation, or creation from scratch), references to queried alignment repositories, and documentation of alignment search engines used in the discovery process.

In the **match phase**, metadata focuses on automated matching execution. This includes matcher configurations (combining algorithm, parameters, and input resources), identification of the matcher used, candidate correspondences generated by the system, evaluation reports assessing their quality, and formal decisions on whether to proceed or return to reconfiguration.

The **align and map phase** documents human refinement and formalisation. Metadata includes alignment plans describing manual adjustments, the final confirmed alignment (set of validated correspondences), and the executable mapping artefact tailored to a specific application.

Finally, the **application phase** captures operational use and evolution. This includes feedback events triggered by application insights or structural data changes, as well as ontology version logs that monitor changes potentially affecting existing mappings.

The project-centric model provides fine-grained metadata coverage not only of technical mapping activities but also of stakeholder involvement, decision-making processes, and reuse evaluation. Compared to earlier lifecycle-based models, it introduces stronger governance and traceability elements. However, similar to OM2R, its terminology and structure remain closely aligned with ontology mapping scenarios, and its applicability across heterogeneous mapping types is not explicitly generalised [5].

Table 3.2 summarises the key metadata fields captured by the project-centric governance model across its lifecycle phases. The full field specifications are provided in Appendix A, Section A.2.

Table 3.2 Key metadata fields captured by the project-centric governance model

Phase	Key Metadata Fields
Stage	Stakeholders, roles, project scope, requirements
Characterise	Discovery report, ontology analysis, source/target selection, mapping decision
Reuse	Reuse evaluation, reuse decision, repository references, search engines used.
Match	Matcher configurations, candidate correspondences, evaluation report, decision
Align and Map	Alignment plan, confirmed alignment, executable mapping artefact
Application	Feedback events, ontology version logs

As shown in Tables 3.1 and 3.2, both lifecycle-based metadata models provide detailed coverage of ontology alignment phases. However, neither model addresses metadata requirements for uplift mappings or entity interlinking specifications. These limitations are examined in detail in Section 3.4.

3.2.2 Artefact-centric and Correspondence-level Models

Unlike lifecycle-based metadata models, artefact-centric approaches focus on representing the mapping artefact itself. These models describe the correspondences between entities and provide limited contextual information about how the mapping was created. Their main objective is to support exchange, interoperability, and tool compatibility rather than documenting governance or lifecycle processes. The following subsections review metadata approaches organised by mapping type.

Ontology Alignment Metadata Approaches

Several metadata approaches have been proposed specifically to annotate and describe ontology alignment. They focus on representing the correspondence between ontology entities. The most prominent of these are SSSOM and the Alignment API format.

SSSOM

SSSOM (Simple Standard for Sharing Ontology Mappings) [14], first published in 2022 and actively maintained since, is a prominent example of an artefact-centric and correspondence-level metadata model. It defines a standardised, machine-readable format for representing mappings between entities. The core structure of SSSOM is built around individual correspondences grouped into mapping sets.

At the correspondence level, SSSOM identifies mapped entities through subject and object identifiers, a predicate describing the relationship, and a mapping justification indicating how the mapping was established. Mapping justifications are expressed using the Semantic Mapping Vocabulary (SEMAPV), which provides controlled terms for describing how a mapping was established, such as lexical matching or manual curation [14]. However, SEMAPV covers only the justification of individual correspondences and does not capture broader lifecycle context such as mapping purpose, design decisions, or stakeholder involvement.

Additional metadata may include confidence values, and provenance information such as the author, creator, reviewer, and mapping date. It also allows documentation of the matching tool used, tool version, similarity measures, preprocessing steps, and other evidence supporting the correspondence.

At the mapping set level, SSSOM provides metadata describing the collection of correspondences as a whole. This includes mapping set identifiers, version information, title, license, and publication date. More recent extensions add elements such as issue tracking references and curation rules to support

maintenance and community feedback, with the data model and tooling continuing to evolve [49]

Although SSSOM is originally developed in the ontology alignment community and aims to extend coverage to other mapping types [14], its core structure assumes pairwise correspondences between entities. This representation copes well with alignment mappings but does not directly accommodate other mapping types, such as uplift mappings or interlinking specifications. As a result, its structure remains closely tied to alignment-type mappings. More specifically, while SSSOM includes a cardinality field to document 1:1, 1:n, n:1, or n:n relationships, it does not support complex correspondence patterns involving algebraic combinations of multiple terms, such as those identified by Scharffe [26]. This is a deliberate design choice to maintain simplicity, meaning that more complex patterns such as attribute transformations or class-by-attribute-value correspondences cannot be fully represented or documented using the current SSSOM model.

Unlike lifecycle-based models, SSSOM does not organise metadata according to explicit mapping phases. Instead, it concentrates on ensuring that each correspondence and mapping set can be clearly identified, exchanged, and reused across tools and infrastructures. Provenance and justification fields are included, but they are not embedded within a structured lifecycle model. As a result, while SSSOM provides strong support for interoperability and exchange of mapping artefacts, it offers limited coverage of stakeholder roles, requirement formulation, reuse decision-making processes, or broader governance activities.

The majority of SSSOM metadata fields are also defined as optional, meaning that in practice mappings can be shared with minimal metadata, which may limit their reusability [14]. Table 3.3 summarises the key metadata fields captured by SSSOM at correspondence and mapping set levels [14]. Only predicate id and mapping justification are required, all other fields are optional. The full SSSOM field specification is provided in Appendix A, Section A.3.

Table 3.3 Key metadata fields in SSSOM.

Level	Key Metadata Fields
Correspondence level	predicate id *(Required)*, mapping justification *(Required)*, subject id, object id, subject label, object label, author id, reviewer id, confidence, mapping date, mapping tool, similarity score.
Mapping set level	mapping set id, license, version, title, description, creator id, mapping date.

Alignment API Format

Alignment API format is another artefact-centric approach that is widely used in ontology alignment [34]. It is commonly adopted in evaluation settings such as the Ontology Alignment Evaluation Initiative (OAEI) for comparing ontology matching systems [16]. The format provides a structured representation of alignments, where each correspondence is expressed through a specified relation and an associated confidence value. As a result, an alignment file is composed of a set of cells, each representing a single correspondence between two entities.

The standard metadata included in an alignment file typically consists of: the URIs or locations of the two ontologies being aligned, the semantic relationship between entities, such as equivalence, a numerical confidence value provided by an alignment algorithm or expert, and the mapping type: such as 1:1, 1:n, n:1, or n:m. The alignment level also captures the method used for generating the alignment and the time taken. These fields reflect the basic format, which prioritises simplicity and tool compatibility over rich documentation.

The format is designed to evolve across different levels of complexity. The basic format is the most commonly used in OAEI evaluations, where entities are strictly defined by their URIs. In contrast, more complex alignments are represented using the Expressive and Declarative Ontology Alignment Language (EDOAL) [35],

which extends the basic format. EDOAL specifically supports complex alignment scenarios by allowing richer semantic expressions.

The **EDOAL** extension captures several advanced features: constructed entities defined through algebraic combinations of other entities, property restrictions such as domain, type, and value constraints, and transformations for property values where different units or encodings are involved.

While the Alignment API format remains a benchmark for tool interoperability, recent efforts often combine it with other standards. For example, SSSOM is frequently recommended for representing simple one-to-one mappings due to its simplicity, whereas EDOAL remains the preferred choice for representing complex mappings.

Despite its expressiveness, the Alignment API primarily focuses on representing correspondences sufficient for tool exchange but provides limited support for lifecycle or contextual metadata. Unlike SSSOM, the format does not include fields for author, date, license, version, or description, metadata that would be essential for governance and reuse. Table 3.4 summarises the key metadata fields captured by the Alignment API format and its EDOAL extension.

Table 3.4 Key metadata fields in the Alignment API format and EDOAL extension [34], [35].

Level	Key Metadata Fields
Alignment level	Ontology 1 URI, Ontology 2 URI, alignment type, alignment level, method, time
Correspondence level	Entity 1 URI, entity 2 URI, relation, confidence measure
EDOAL extension	Constructed entities, property restrictions, property transformations, linkkeys

The full field specifications are provided in Appendix A, Section A.4.

Uplift Mapping Metadata Approaches

In the context of data transformation mappings (uplift mappings), particularly RML-based knowledge graph construction, several efforts have focused on improving the documentation and metadata annotation of mappings. These efforts have evolved over time from machine-generated metadata toward approaches that also aim to improve the understanding and usability of mappings.

An early contribution in this direction is the work proposed by **Dimou et al. [4]**, which introduces an approach for automating the generation of metadata and provenance within linked data generation and publishing workflows. Metadata is generated across four stages of the workflow: mapping definition, data source retrieval, RDF generation, and publication. It also supports annotating mappings at different levels of granularity, ranging from dataset-level to triple-level and RDF term-level metadata.

Rather than defining a fixed set of metadata fields, the approach generates metadata dynamically using PROV-O [11], and DCAT [12]. This covers provenance metadata describing activities, entities, and agents, dataset-level metadata such as identifiers, descriptions, and publishers, and mapping-related metadata associating information with mapping documents and their components such as Triples Maps.

A key strength of this approach is that metadata generation is integrated into the RML-based data transformation workflow. As a result, metadata is produced automatically during execution, enabling machine-readable, consistent, and reproducible metadata without requiring manual annotation. This provides strong support for traceability and reproducibility of mapping execution.

However, the captured metadata focuses on provenance and execution-related aspects, describing how mappings are executed rather than why they are designed. Contextual information such as mapping purpose, design decisions, assumptions, stakeholder involvement, and evaluation processes is absent from this model. The approach is also tightly coupled to RML workflows and does not discuss its

applicability to other types of mappings, such as ontology alignment or interlinking.

More recent work has shifted from the automatic generation of metadata toward improving the understandability of mapping specifications. **RMLdoc** is a tool designed to generate human-readable documentation for mappings expressed in R2RML, RML, and YARRRML [50]. RMLdoc processes mapping documents and extracts relevant metadata and structural components, then generates human-readable documentation accompanied by visual diagrams. In terms of metadata coverage, the tool captures metadata at the mapping document level, such as title, version, authors, description, creation date, and license. In addition, it extracts structural information about the mapping itself, including logical sources, subject maps, predicate-object maps, and join conditions. These elements are presented in a human-readable format to support better understanding of the mappings.

This approach addresses an important limitation of existing mappings, namely the difficulty of interpreting and understanding complex mapping documents. By transforming these complex specifications into visual and accessible documentation, RMLdoc supports knowledge engineers in understanding and analysing mappings more easily.

However, similar to previously discussed approaches, the focus of RMLdoc remains limited in scope. The captured metadata is primarily intended for documentation and readability purposes, rather than for alignment with FAIR principles or supporting reuse. The approach is also designed specifically for RML mappings and does not explicitly capture contextual or lifecycle metadata, such as mapping purpose, design decisions, or reuse considerations.

A significant recent contribution to the RML-based mappings is the **RML ontology** [51], which redesigns RML as a modular OWL ontology. The ontology provides a formal vocabulary for describing RML mapping components such as logical sources, subject maps, predicate-object maps, join conditions, reference formulations, term maps, graph maps, and language maps. However, while this

contribution advances the machine-readable mapping specification, it focuses on the structural components of the mapping rules rather than lifecycle metadata such as mapping purpose, design decisions, or stakeholder involvement.

Both approaches (RMLdoc and RML ontology) contribute to the documentation and formalisation of RML mappings. The main difference between them is that RMLdoc focuses on generating human-readable documentation while RML ontology provides formal machine-readable vocabulary for describing the structural components of RML. However, neither captures lifecycle metadata nor addresses mapping types beyond RML-based uplift scenarios. Table 3.5 summarises the key metadata fields captured by these RML approaches discussed in this section.

Table 3.5 Key metadata fields captured by RML approaches

Approach	level	Key Metadata Fields
Dimou et al. [4]	Dataset / triple / term level	Provenance activities, entities, agents, dataset identifiers, descriptions, publishers, access information, mapping document references, Triples Map associations
RMLdoc [50].	Document level	Title, version, authors, description, creation date, license, logical sources, subject maps, predicate-object maps, join conditions
RML Ontology [51]	Component level	Logical sources, subject maps, predicate-object maps, join conditions, reference formulations, term maps, graph maps, language maps

The full field specifications are provided in Appendix A, Section A.5. As shown in Table 3.5, all three approaches capture metadata at the execution or document level without reference to the broader lifecycle context in which mappings are created, and none addresses ontology alignment or entity interlinking scenarios.

Entity Interlinking Metadata Approaches

Unlike ontology alignment and uplift mapping, entity interlinking specifications, to the best of the author's knowledge, lack dedicated metadata models. Tools such as Silk [31] and LIMES [32] provide link specifications that define the conditions

for generating links between instances across datasets. However, these specifications do not include structured metadata describing the purpose, provenance, or lifecycle context of the interlinking process. Link quality and evaluation results are not captured in a standardised format, and no community standard equivalent to SSSOM exists for interlinking metadata, although recent initiatives such as the RDA FAIR Mappings Working Group have begun to address this gap. This absence represents a notable gap in the current landscape and further supports the need for a unified metadata model applicable across all declarative mapping types.

As shown across the three mapping types reviewed in this section, artefact-centric approaches share a common limitation: metadata is captured at the correspondence, document, or execution level without reference to the broader lifecycle context in which mappings are created, and no unified approach addresses all three declarative mapping types. These limitations are examined in detail in Section 3.4.

3.2.3 FAIR-Oriented Metadata Frameworks

Mappings are no longer treated as standalone technical artefacts but as reusable semantic artefacts that require rich and contextual metadata to describe them. Clear evidence of this can be seen in recent efforts from the knowledge graph and semantic web community, which have increasingly focused on aligning mapping metadata with the FAIR principles. These efforts are closely connected to broader initiatives such as the European Open Science Cloud (EOSC), which provides an environment for sharing and reusing research data and digital artefacts across domains. Within EOSC, mappings play a key role in enabling semantic interoperability between heterogeneous data sources, which emphasises the need to treat them as first-class artefacts that should be findable, accessible, interoperable, and reusable.

In the past few years, several initiatives have emerged that aim to make mappings FAIR semantic artefacts. These initiatives are closely related but operate at different levels. FAIR-IMPACT focuses on defining mapping practices and developing guidelines for mappings in order to align with FAIR principles [48] [9]. Building on and extending these efforts, the Research Data Alliance (RDA) FAIR Mappings Working Group works with different communities toward proposing common guidelines for making different types of mappings FAIR [18]. It aims to move beyond existing specifications by promoting a common, machine-readable and actionable representation of mappings. The RDA FAIR Mappings Working Group explicitly builds on FAIR-IMPACT and extends its work beyond Europe, while also considering infrastructure developments such as those provided by FAIRCORE4EOSC [18].

FAIRCORE4EOSC focuses on developing the infrastructure required to support FAIR Digital Objects within EOSC [52], including mechanisms for registering mappings through persistent identifiers and managing their metadata. In this context, mappings are treated as digital artefacts that can be identified, described, and accessed across systems.

Together, these initiatives highlight the increasing importance of treating mappings as first-class artefacts and demonstrate a progression from defining guidelines, to community alignment, to infrastructure support. Each of these initiatives is discussed in more detail in the following subsections, with a focus on their approaches to mapping metadata and their limitations.

FAIR-IMPACT

One of these initiatives is FAIR-IMPACT, which has been actively working on proposing recommendations and guidelines for defining mapping practices and making mappings FAIR artefacts. Their work highlights the importance of treating mappings as first-class artefacts, as they are central to promoting interoperability between information systems. It also emphasises the need for persistent identifiers to make mappings findable, rich metadata to describe them, and accessible

repositories to store them, a particularly important contribution given that mappings are currently very difficult to find and lack the contextual information required to support reuse.

As described in Section 3.1.3, the initiative also proposes a set of phases as a framework to organise mapping practices. However, these phases are not explicitly linked to a structured metadata model. FAIR-IMPACT also provides a set of metadata recommendations that extend models such as SSSOM by introducing additional metadata elements required to support FAIR principles [9], [40].

FAIR-IMPACT compiled a list of 19 metadata elements identified as essential for describing mappings and crosswalks in a comprehensive manner [9], [40]. These metadata elements are aligned with the SSSOM model with additional metadata addressing missing information and gaps. Key additions include the context of the mapping, which captures the purpose and use case of the mapping. This is not explicitly considered in the SSSOM model and is identified as essential for reuse. Representing source and target artefact names is also identified as essential to support discoverability and is not explicitly covered in the SSSOM model. FAIR-IMPACT also highlights that most of these metadata elements are defined as optional in SSSOM; however, in order to fully support findability and reusability, these elements may need to be treated as mandatory.

In terms of metadata coverage level, FAIR-IMPACT considers both mapping-level and crosswalk-level metadata. A mapping refers to an individual correspondence between source and target elements (a single mapping assertion), for example mapping Person to *foaf:Person*, whereas a crosswalk represents a collection of mappings defined for a specific integration task (a mapping set), for example mapping a Dublin City dataset to a national statistics dataset. This reflects an understanding that metadata can be captured at different levels of granularity to support discovery and reuse. However, while these levels are acknowledged, the framework does not fully distinguish between different levels of abstraction, such

as project-level, mapping-level, and correspondence-level metadata in a unified and structured way.

Beyond metadata recommendations, FAIR-IMPACT recognises the need to support different types of mappings, including more complex mappings where a single source element may map to multiple target elements. The initiative proposes extending the SSSOM model by defining a core structure based on source, target, and predicate. However, this extension remains at a conceptual level, and how to annotate and represent complex mappings across different mapping types is not yet fully defined.

While FAIR-IMPACT provides valuable recommendations and guidelines, its outputs remain largely conceptual and do not define a concrete, unified metadata model that can be directly implemented across different mapping types. Questions also remain regarding the practicality of such metadata annotation, including whether it may produce large mapping files that could introduce challenges in terms of storage, efficiency, and query processing, or whether contextual metadata should be limited to a certain level of granularity.

RDA FAIR Mappings

The RDA FAIR Mappings Working Group explicitly builds on previous initiatives, particularly FAIR-IMPACT, and extends this work beyond Europe, while also considering infrastructure developments such as those provided by FAIRCORE4EOSC.

The goal of this group is to identify a set of technical recommendations for designing FAIR mappings, a methodology covering different steps of the mapping lifecycle, a classification of mapping types, and a set of data models and associated metadata for sharing mappings. The group also proposes a conceptual model that distinguishes between mapping specification and mapping execution, and introduces key concepts and terminology such as mapping relationships, mapping activities, and mapping sets.

In terms of metadata coverage, the RDA FAIR Mappings Working Group builds on SSSOM and extends it by proposing additional metadata elements. These include provenance metadata (e.g., author, creator, reviewer, and publication date), descriptive metadata (e.g., mapping method, justification, and context), and source and target metadata (e.g., identifiers, names, and versions of semantic artefacts). The group also proposes a minimal FAIR mapping schema that aims to provide a standardised way to describe mapping specifications across different mapping formats.

Despite these contributions, the work of the RDA FAIR Mappings Working Group remains ongoing. While it provides important steps toward proposing standards and machine-readable representations, it does not yet define a fully unified and stable metadata model that can be consistently applied across all mapping types and lifecycle phases. Although lifecycle steps are acknowledged, they are not explicitly structured within a comprehensive lifecycle-based metadata model. These limitations, shared across all approaches reviewed in this section, are examined in detail in Section 3.4.

3.3 AI and LLM Based Mapping Generation Approaches

Recent advances in large language models have generated significant interest in their application to knowledge graph construction and declarative mapping generation. LLMs offer an alternative to both manually authored specifications or on fully automated systems, enabling generation mapping-related tasks including generating RDF triples, aligning ontologies, interlinking datasets, and producing mapping rules from natural language descriptions [19], [53], [54], [55]. This capability offers knowledge engineers and data practitioners new opportunities to reduce the manual effort involved in both knowledge graph construction and mapping creation [56], [57]. However, it also introduces challenges related to semantic precision, reproducibility, and contextual grounding that are directly relevant to the research presented in this thesis.

RML uplift mapping generation

The use of LLMs for RML uplift mapping generation has received growing attention within the knowledge graph construction community. To the best of the author's knowledge, **Hofer et al.** conducted the first systematic evaluation of LLM capabilities for generating RML mapping files in Turtle format using a zero-shot experimental setup, in which five commercial models were tasked with mapping a JSON subset of the IMDB dataset to the DBpedia Movie ontology [19]. Their evaluation found that syntactic errors were common across all models, although most were capable of repairing them through a secondary repair prompt. A more critical finding emerged when valid mappings were executed, the majority failed to produce any RDF triples, with incorrect iterator definitions and reference conditions identified as the primary cause. Models also consistently confused semantically similar ontology properties, demonstrating that the core challenge in LLM-based RML generation is not syntactic but semantic.

The prompt used in this study embedded informal rules specifying property selection strategy, datatype handling, namespace conventions, and domain constraints rules that correspond directly to the structured metadata fields proposed in this thesis, where mapping purpose, domain assumptions, design decisions, and technical requirements are captured formally across the mapping lifecycle. This suggests that lifecycle-based metadata could serve as a reusable and structured foundation for the kind of contextual guidance that was expressed informally in their work. Figures 3.5 and 3.6 illustrate the mapping generation and repair prompts used in Hofer et al.

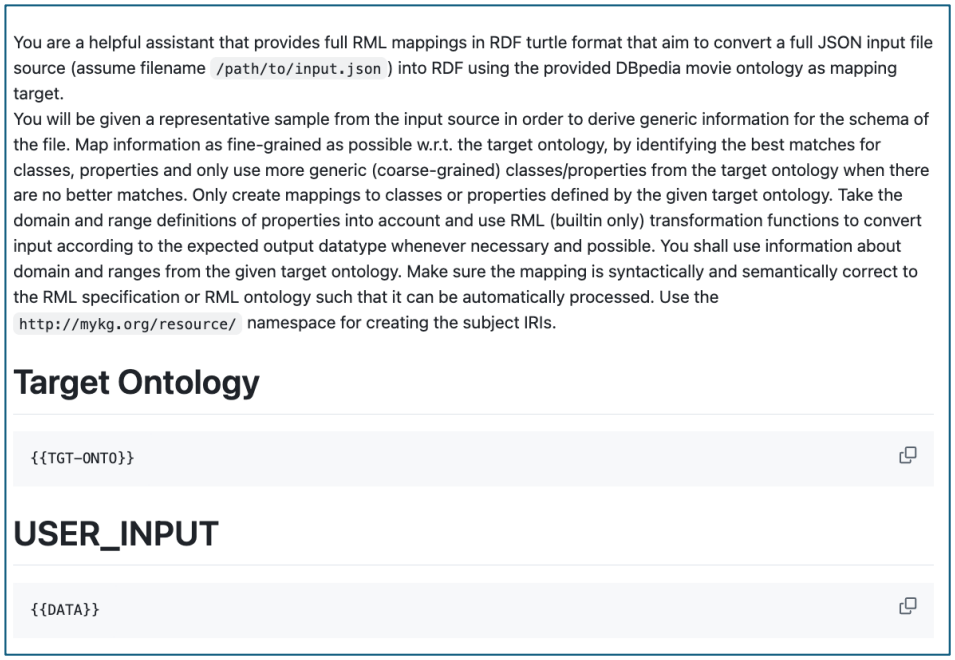


Figure 3.5 Mapping generation prompt used by Hofer et al.

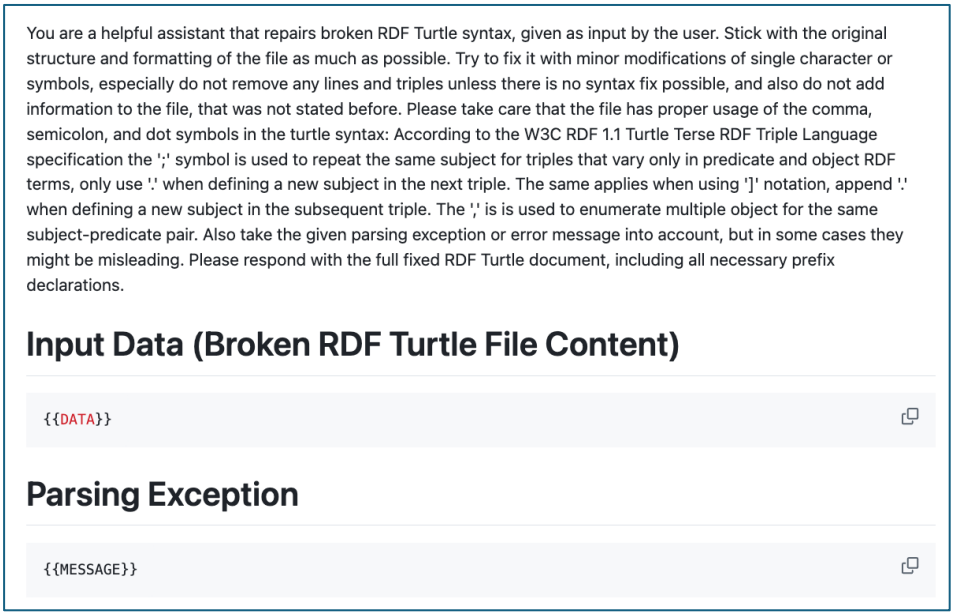


Figure 3.6 Syntax repair prompt used by Hofer et al.

Schmidt et al. developed **MYAM+R**, a pipeline for LLM-assisted RML and YARRRML generation in a manufacturing context, evaluated across real-world Bosch datasets [55]. The pipeline uses automated ontology reduction to address token limitations when dealing with large ontologies, and prompt enhancements

including mapping templates and few-shot examples to improve syntactic quality. However, their prompt configurations are limited to syntax guidance and role assignment, providing no structured representation of mapping purpose, domain assumptions, design decisions, or quality requirements. Figure 3.7 illustrates the zero-shot prompt configuration used in their work.

```
Role:
You are a Senior Manufacturing Knowledge Expert tasked with creating
a YARRRML mapping that aligns the provided manufacturing dataset with
the specified ontology.

Input:
You will receive:

    A dataset in CSV format.
    An ontology in Turtle format.

Instructions:

    Map Column Headers: For each column header in the dataset,
    identify and map it to the corresponding elements defined in the
    ontology. If a column header does not directly correspond to an
    ontology element, consider the context of the data to find the
    most appropriate mapping.
    Create Mapping Blocks: For each ontology class that is
    represented in the dataset, create a separate mapping block.
    Ensure that all relevant properties and relationships for that
    class are included in the mapping.
    Utilize Namespaces and Prefixes: Use the namespaces and prefixes
    from both the given ontology and any relevant upper ontologies
    for the mapping.
    Response Format: Respond solely with the YARRRML mapping; do not
    include any natural language explanations or commentary.
    Adhere to YARRRML Syntax: Ensure adherence to the YARRRML syntax
    throughout your response.

```

Figure 3.7 Zero-shot prompt configuration used in MYAM+R by Schmidt et al.

Frey et al. evaluated LLM capabilities for RDF Turtle generation and SPARQL query creation across fifteen model versions using the LLM-KG-Bench framework [58]. Their main finding was that while syntax errors could be corrected through multi-shot repair prompts, semantic correctness remained a persistent challenge, particularly for SPARQL generation where most models produced empty result sets. Their benchmark tasks used prompts ranging from simple task descriptions to detailed step-by-step instructions specifying expert roles, namespace conventions, and output format requirements. As with the previous approaches reviewed in this section, none incorporated structured representations of mapping purpose, domain

assumptions, or design decisions. Figures 3.8 and 3.9 illustrate a simple and the most contextually rich prompt used in their benchmark respectively.

> **Prompt 3:** Create a knowledge graph in turtle format that consists of n different objects of type `foaf:person`. Each should have at least 2 and at most $n-1$ connections to other persons via `foaf:knows`. Give no extra text.

Figure 3.8 Simple task prompt used by Frey et al.[59]

> **Prompt 5:** As a Linked Data expert and knowledge engineer please convert the 3d printer specification given in the bottom into an RDF turtle formatted Knowledge Graph. The main subject in this Knowledge Graph should be the Printer itself using the https://data.semper-ki.org/resources/$encoded_label$ whereas $encoded_label$ label refers to the name of the printer in which whitespaces have been replaced with underscores. Step by step attach the following statements to that node.
> 1) Attach https://purl.org/tema/051993 as rdf:type
> 2) Attach the manufacturer via an object property using the schema.org vocabulary. For the subject identifier of the manufacturer use the same approach as for the printer subject and assign it the rdfs:label without a language tag.
> 3) Attach all printable material as well all support materials the printer can use via object properties. For the property identifiers use https://data.semper-ki.org/properties/printMaterial respectively https://data.semper-ki.org/properties/supportMaterial and for material identifiers use the https://data.semper-ki.org/resources/materials/ namespace.
> 4) Attach the dimensions of the printer as 3 separate values using the schema.org vocabulary for the property identifiers of the dimensions, but for the values use qudt:QuantityValue objects from the QUDT ontologies family to have the numerical value (via qudt:numericValue property and typed as xsd:double) and the appropriate QUDT unit of measurement identifier (via qudt:hasUnit property) separately. Do not convert the dimensions into another unit.
> Follow best practices of Linked Data and common standard vocabularies as basis for the modeling unless specified differently in the requirements above. Do not extract any other properties or values from the specification than the ones I mentioned but validate your output step by step, to check whether all used prefixes are included and the given requirements as well as the grammar of the RDF turtle serialization were respected. Only return the turtle format, no additional text. [... followed by the fact sheet plaintext excerpt]

Figure 3.9 Most contextually rich task prompt (T5) used by Frey et al.

Randles and O'Sullivan explored the integration of ChatGPT into uplift mapping workflows through the **R2[RML]-ChatGPT** Framework. Their framework uses LLMs to assess the semantic and syntactic correctness of RML mappings during linked data publication workflows, establishing that LLMs can contribute meaningfully to mapping quality assessment when properly integrated into structured workflows [57]. Their predefined prompt templates insert only individual ontology concept names to retrieve quality-related information from

ChatGPT, for example: *"Can you provide me with key information related to the '[concept_name]' used in RDF/OWL Technology?"*, providing no structured representation of mapping purpose, design decisions, or lifecycle context beyond the concept name itself.

Ontology Alignment generation

In the ontology alignment community, researchers have demonstrated the potential of LLMs as matching assistants. **Babaei Giglou et al.** proposed **LLMs4OM**, a framework for ontology matching using retrieval-augmented matching modules with zero-shot prompting [54]. The framework evaluates three concept representations which are concept name only, concept with parent classes, and concept with child classes, to assess how different levels of ontology structure affect matching quality. Their evaluation across twenty OAEI datasets suggests that LLMs can match and in some cases surpass traditional ontology matching systems. Figure 3.10 shows their prompt templates provide only ontology structural information such as concept names and hierarchical relationships, with no representation of alignment purpose, domain assumptions, design decisions, or quality requirements.

Prompt template for CP representation.

Classify if two concepts refer to the same real world entity or not (answer only yes or no).
First concept:
$\{C_s\}$
Parents: $\{CP\}$
Second concept:
$\{C_t\}$
Parents: $\{CP\}$
Answer:

An example:

Classify if two concepts refer to the same real world entity or not (answer only yes or no).
First concept:
cardiovascular system
Parents: organ system
Second concept:
vascular endothelium
Parents: endothelium, blood vessel tissue
Answer:

Figure 3.10 Prompt template and example used by Babaei Giglou et al.

Jiménez-Ruiz et al. explored the use of LLMs as validation oracles within the LogMap alignment system, restricting LLM consultation to uncertain correspondences and evaluating thirty LLM configurations across nine OAEI matching tasks, demonstrating that selective LLM consultation can effectively support alignment validation [59]. As illustrated in Figure 3.11, even their most contextually rich prompts are limited to ontology structural information including synonyms and full hierarchical lineage, with no structured representation of alignment purpose, domain assumptions, or design decisions.

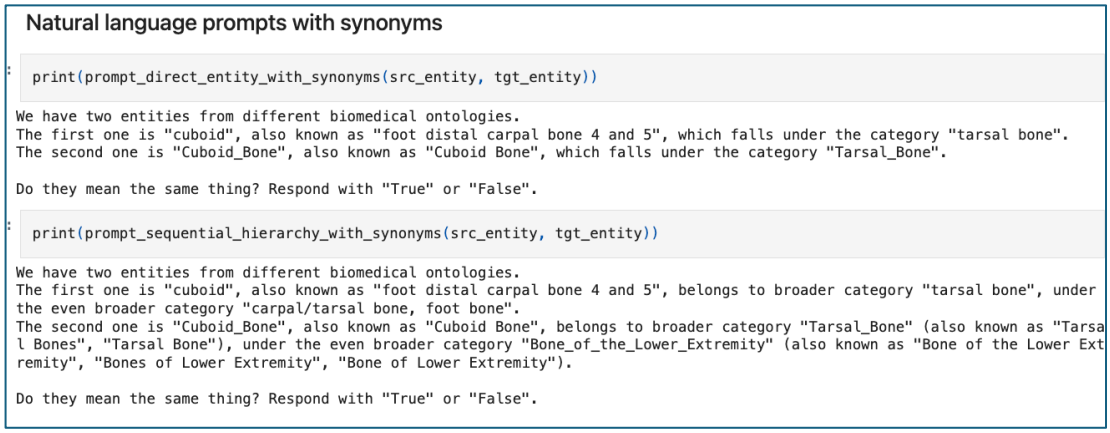


Figure 3.11 Simple and most contextually rich prompt variants used by Jimenez Ruiz et al.

Taboada et al. introduced **MILA**, a retrieve-identify-prompt pipeline embedded within a prioritised depth-first search strategy that invokes the LLM only for borderline correspondences [60]. Their approach achieved the best F-measure in five of seven unsupervised tasks in the OAEI 2024 evaluation, outperforming state-of-the-art systems by up to 17%. As shown in Figure 3.12, their prompt template provides only ontology names and concept labels, with no structured representation of alignment purpose, domain assumptions, or design decisions.

You are a helpful expert in ontology matching, which involves determining equivalence correspondences between concepts from different ontologies. The source ontology is called [*src_onto_name*] and the target ontology is called [*tgt_onto_name*].
Classify whether the following concepts are equivalent:
Source concept: [*source_entity*]
Target concept: [*target_entity*]
If so, answer 'Yes', without adding any type of explanation. Otherwise, answer 'No'.

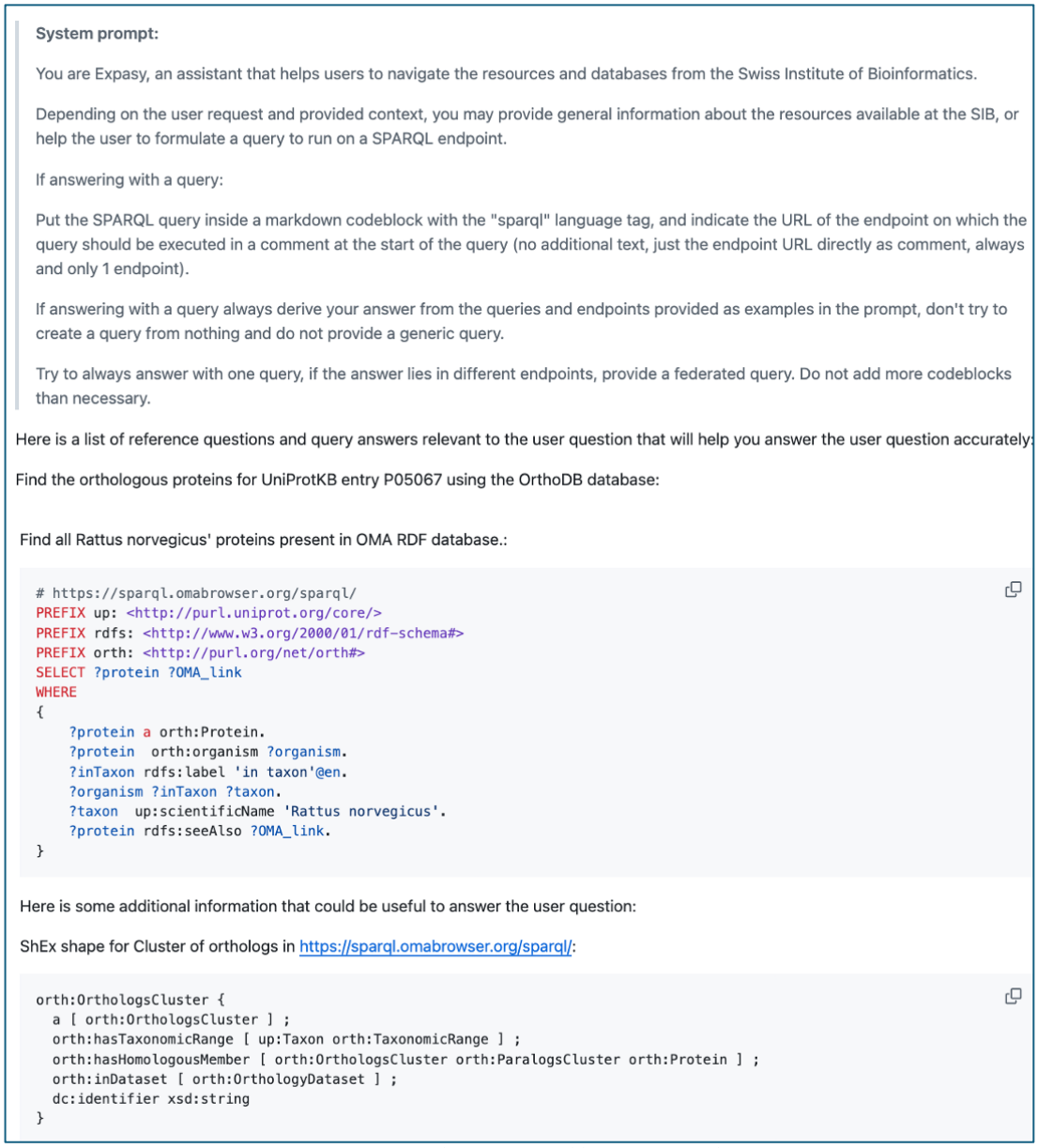
Figure 3.12 Prompt template used by Taboada et al. in MILA for ontology concept matching

A consistent finding across these three approaches is that constraining when and how the LLM is consulted significantly improves both accuracy and efficiency. While the prompt designs vary in their level of detail from simple concept names in Babaei Giglou et al. to hierarchical lineage in Jiménez-Ruiz et al. and expert role assignment in Taboada et al., none incorporates structured lifecycle context such as alignment purpose, domain assumptions, or design decisions.

Interlinking generation

In the context of entity interlinking and SPARQL query generation, LLM based approaches have demonstrated both promise and limitation**. Emonet et al.** proposed a retrieval-augmented generation system for translating natural language questions into federated SPARQL queries over a bioinformatics knowledge graph [53]. Their system uses schema metadata and query examples as contextual input to the LLM. An iterative validation and correction loop is then applied to refine the generated queries. Their use of structured schema metadata as contextual input to improve generation quality directly parallels the lifecycle metadata approach investigated in this thesis. However, their context remains limited to schema-level information rather than project-level lifecycle knowledge. As illustrated in Figure 3.13, their prompt includes a system role description, retrieved example question-query pairs, and ShEx class schemas as contextual input, all representing structured technical information that improves generation quality, yet none capturing mapping purpose, domain assumptions, or design decisions.

Among the approaches reviewed in this section, **Emonet et al.** represent the closest parallel to the approach proposed in this thesis. Both use structured contextual input to guide the LLM towards more complete and correct outputs. However, their structured context is limited to technical schema information. The lifecycle-based metadata model proposed in this thesis goes further by providing project-level contextual guidance including mapping purpose, domain assumptions, design decisions, and quality requirements. This offers a reusable and formally structured foundation for LLM-assisted mapping generation, avoiding ad hoc prompt construction.



System prompt:

You are Expasy, an assistant that helps users to navigate the resources and databases from the Swiss Institute of Bioinformatics.

Depending on the user request and provided context, you may provide general information about the resources available at the SIB, or help the user to formulate a query to run on a SPARQL endpoint.

If answering with a query:

Put the SPARQL query inside a markdown codeblock with the "sparql" language tag, and indicate the URL of the endpoint on which the query should be executed in a comment at the start of the query (no additional text, just the endpoint URL directly as comment, always and only 1 endpoint).

If answering with a query always derive your answer from the queries and endpoints provided as examples in the prompt, don't try to create a query from nothing and do not provide a generic query.

Try to always answer with one query, if the answer lies in different endpoints, provide a federated query. Do not add more codeblocks than necessary.

Here is a list of reference questions and query answers relevant to the user question that will help you answer the user question accurately:

Find the orthologous proteins for UniProtKB entry P05067 using the OrthoDB database:

Find all Rattus norvegicus' proteins present in OMA RDF database.:

```
# https://sparql.omabrowser.org/sparql/
PREFIX up: <http://purl.uniprot.org/core/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX orth: <http://purl.org/net/orth#>
SELECT ?protein ?OMA_link
WHERE
{
    ?protein a orth:Protein.
    ?protein  orth:organism ?organism.
    ?inTaxon rdfs:label 'in taxon'@en.
    ?organism ?inTaxon ?taxon.
    ?taxon  up:scientificName 'Rattus norvegicus'.
    ?protein rdfs:seeAlso ?OMA_link.
}
```

Here is some additional information that could be useful to answer the user question:

ShEx shape for Cluster of orthologs in https://sparql.omabrowser.org/sparql/:

```
orth:OrthologsCluster {
  a [ orth:OrthologsCluster ] ;
  orth:hasTaxonomicRange [ up:Taxon orth:TaxonomicRange ] ;
  orth:hasHomologousMember [ orth:OrthologsCluster orth:ParalogsCluster orth:Protein ] ;
  orth:inDataset [ orth:OrthologyDataset ] ;
  dc:identifier xsd:string
}
```

Figure 3.13 Example prompt used by Emonet et al.

Ihtassine et al. have proposed **NL2LS**, a framework for generating declarative link specifications from natural language for LIMES and Silk frameworks, evaluating zero-shot, one-shot, and supervised fine-tuning approaches [61]. Their results show that zero-shot performance was consistently poor across all models, with fine-tuning required to achieve reliable generation. As illustrated in Figures 3.14 and 3.15, their prompt configurations range from a simple task instruction in the zero-shot setting to a refined prompt with expert role assignment and syntax

rules. However, neither incorporates domain context, mapping purpose, or lifecycle metadata which highlights precisely the gap that the lifecycle-based metadata model proposed in this thesis addresses, where such project-level contextual knowledge is formalised as structured and reusable input.

```
def build_zero_shot_prompt(example):
    prompt = (
        "Task: Translate the following natural-language description into a Link
        Specification.\n"
        "Output rules:\n"
        " - Output ONLY the Link Specification (single line).\n"
        " - Do NOT add explanations or extra text.\n\n"
        f"Text:\n{example['nl']}\n"
        "Answer:"
    )
    return {"prompt": prompt, "ls": example["ls"]}
```

Figure 3.14 Simple zero-shot prompt used by Ihtassine et al. in NL2LS

```
def gpt_4o_parse_to_ls(verbalized_text):
    """
    Uses GPT-4o to parse a verbalized description into a Link Specification (LS).
    :param verbalized_text: The natural language description of the LS.
    :return: The parsed LS in machine-readable format.
    """
    # Define the refined prompt for the LLM
    prompt = [
        {"role": "system", "content": "You are an expert in translating verbalized
        logic into machine-readable format."},
        {"role": "user", "content": f"""
        You are an expert in knowledge graphs and link specifications. Your task is
        to translate a natural language description of link specifications into a
        machine-readable format.

        The input description is as follows:
        "{verbalized_text}"

        The output should follow these rules:
        - Use the format: <metric>(<entity1>, <entity2>) | <threshold>
        - Logical operators like "or" and "and" are binary operators with two
        operands separated by a comma and should be written as OR() and AND().
        operators with more than 2 operands must by neasted
        - Conditions should be grouped and nested based on their logical hierarchy in
        the input description.
        - Do not flatten nested logical structures.
        - The final output should reflect the exact logical grouping implied in the
        input.

        Translate the given description into the desired format with the correct
        nested structure.
        Provide the answer only in one line.
        """}
    ]

    # Query the LLM using the query_llm function
    response = query_llm(prompt)

    # Extract and return the result
    return response["response"]
```

Figure 3.15 Refined prompt used by Ihtassine et al. in NL2LS

Discussion

Despite the progress demonstrated across the approaches reviewed in this section, several challenges persist. A systematic literature review of LLM-based ontology engineering confirmed that LLM-supported tasks are concentrated in early development phases, with lifecycle coverage across maintenance, reuse, and quality evaluation remaining largely absent [62],under review. The review also found that more semantically rich and contextually grounded prompts consistently lead to more correct outputs. However, no unified structured template exists to guide practitioners in constructing such prompts. This is directly observable across the approaches reviewed in this section, where metadata fields proposed in this thesis such as stakeholder roles, mapping purpose, domain assumptions, design decisions, and quality requirements appear in fragmented form across different prompt configurations, never unified or formalised.

Agrawal et al. demonstrated that LLMs frequently hallucinate in knowledge graph tasks and that grounding the LLM with structured contextual knowledge significantly reduces this problem [63]. **Freund et al.** proposed **ReMap**, an algorithmic approach to RML mapping generation, demonstrating that a structured example-based approach outperforms unguided LLM generation by 31% on RML core test cases [64]. These findings confirm that the core challenge across all reviewed approaches is contextual grounding. The metadata fields needed to guide generation are present in fragments but never captured formally, structured consistently, or provided as reusable lifecycle-based input. **Also, it is important to note that the lifecycle-based metadata model proposed in this thesis is not intended to replace existing prompt engineering techniques or mapping tools, but rather to complement them by providing practitioners with a structured and reusable contextual foundation that can be adapted as a guided template to different mapping types and lifecycle stages.**

Table 3.7 summarises the prompt context and output annotation across all eleven approaches reviewed in this section. The pattern is consistent, lifecycle metadata fields appear in fragments but are never unified, formalised, or provided as reusable input. This gap was first identified in earlier work examining metadata requirements for declarative mappings [65], [66], and is directly addressed in this thesis by proposing a lifecycle-based metadata model that captures these fields formally across all three mapping types.

Table 3.6 Overview of prompt context and output annotation across LLM-based mapping generation approaches

Approach	Mapping Type	Prompt Context	MMV Fields in Fragments	Metadata Attached to Output
Hofer et al. [19]	RML uplift	Ontology + task description + informal domain rules	mmv:hasDesignDecision, mmv:hasDomainAssumption, mmv:hasTechnicalRequirement	No
Schmidt et al. [55]	RML/YARRRML	Task description + role + syntax template	mmv:hasStakeholder only	No
Frey et al. [58]	RDF/SPARQL	Task description + output format	None	No
Randles & O'Sullivan [57]	RML uplift	Task description + concept name	None	No
Babaei Giglou et al. [54]	Alignment	Task description + concept names + hierarchy	None	No
Jiménez-Ruiz et al. [59].	Alignment	Task description + names + synonyms + lineage	None	No
Taboada et al. [60]	Alignment	Task description + ontology names + concept labels	None	No
Emonet et al. [53]	SPARQL	Task description + schema metadata + query examples + ShEx	Fragment of mmv:hasTechnicalRequirement	No
Ihtassine et al. [61]	Interlinking	Task description + output format	None	No
Freund et al. [64].	RML uplift	Example input-output pairs algorithmic	None	No
This thesis	All three	Task description + All MMV fields (formally structured and unified)	All MMV fields (formally structured and unified)	Yes

3.4 Analysis of Limitations and Open Challenges in the State of the Art

The review of the lifecycle frameworks, metadata models, and FAIR-oriented initiatives shows how the structure and documentation of declarative mappings have evolved over the years. It also reveals the growing recognition within the community of the importance of treating declarative mappings as first-class artefacts, as well as aligning them with FAIR principles and best practices. However, it also highlights several limitations and open challenges that remain unsolved across the current state of the art.

One of the main limitations that can be recognised immediately is the fragmentation of metadata approaches. Existing metadata solutions tend to focus on specific types of mappings. Early lifecycle-based model proposals such as OM2R provide rich contextual and governance metadata, but they were primarily proposed for ontology alignment. In contrast, artefact-centric approaches such as SSSOM and the Alignment API focus on describing mappings at the correspondence level, but with limited contextual information, and are also mainly designed for ontology alignment scenarios. Similarly, approaches in data transformation mappings (uplift mappings), such as RMLdoc, focus on provenance, execution, and understanding of the mapping document, but do not capture the broader lifecycle context. As a result, metadata remains distributed, and different approaches suggest different metadata for mappings, even though they share common development practices and technologies, despite differences in structure or representation. This fragmentation is further evident at the correspondence level, where patterns such as those identified by Scharffe and Fensel [26] appear across all three mapping types, yet no existing metadata approach captures them in a unified way.

A second challenge is the limited support for mapping reuse. Declarative mappings often require knowledge engineers and domain experts to be involved, and they also require time and resources to develop. Although many approaches

acknowledge the importance of reuse, explicit support for reuse decision-making is often missing. Existing models do not provide sufficient metadata to allow users to assess whether a mapping is suitable for a new context or can be used as a starting point for a new mapping project, where the level of reuse may vary. Important information such as the purpose of the mapping, domain, assumptions, technical requirements, people involved, constraints, design decision justification, and evaluation and quality results is often absent or only partially captured. This limitation is also reflected in FAIR-oriented initiatives, where reuse is emphasised, but metadata related to quality and evaluation is still missing or not fully captured, which can affect the trustworthiness and reusability of mappings.

Another important limitation relates to the representation of metadata across different levels of granularity. While some approaches distinguish between mapping-level (correspondence level), which is the finer level, and crosswalk-level (mapping document level), there is no unified framework that clearly integrates these levels, distinguishes between them, or assesses to what level it is feasible to annotate mappings in terms of storage and query processing time. This is an important aspect to consider, as these proposals aim to make mappings findable and reusable, which inherently involves querying and processing. As a result, it remains unclear how metadata should be structured across project-level, mapping-level, and finer levels, and whether high-level metadata description alone is sufficient to support initial reuse decisions and trustworthiness before further correspondence-level inspection.

These limitations collectively point toward a deeper issue: the lack of a unified lifecycle-based metadata model that can be applied consistently across different mapping types. As discussed previously, declarative mappings and mapping activities in the field often share common practices. Having a common structure and framework to organise the mapping process would contribute to structuring both the development process and the metadata capturing across different stages, starting from domain and scope definition, stakeholder involvement, and design decisions, to development processes, tools used, and ongoing maintenance and

testing. Early lifecycle framework proposals such as OISIN and OM2R provide structured representations of the mapping process, but they remain focused on ontology alignment. Recent initiatives also highlight the importance of having a common framework for structuring mapping activities, but their proposed metadata remains independent of such lifecycle structures.

In addition, there is still a gap between conceptual FAIR recommendations and practical implementation. Initiatives such as FAIR-IMPACT and the RDA FAIR Mappings Working Group provide valuable guidelines and identify important metadata requirements. However, these efforts are relatively recent and remain largely conceptual or infrastructural. They do not yet define a concrete, unified metadata model that can be directly implemented and applied to support mapping documentation, discovery, and reuse across different mapping types and lifecycle stages.

In parallel to these developments, emerging LLM-based approaches for mapping generation are reviewed in detail in Section 3.3. These approaches introduce new opportunities for automating mapping creation across all three mapping types. However, current LLM-based systems operate without structured guided templates or lifecycle metadata. They rely instead on ad hoc natural language prompts that provide no representation of mapping purpose, domain assumptions, design decisions, or quality requirements. The lifecycle-based metadata model proposed in this thesis addresses this gap directly. It captures what happens during the development of declarative mappings across different lifecycle stages. This offers a structured and reusable foundation that can serve as a guided template for LLM-assisted generation. It also provides a metadata annotation model for understanding and interpreting the generated output.

Based on the review presented in this chapter, six key gaps can be identified:

Gap 1: No unified metadata model exists that addresses all three declarative mapping types including ontology alignment, uplift mappings, and entity interlinking in a single framework.

Gap 2: Existing metadata models either focus on artefact exchange without lifecycle context, or propose lifecycle structures limited to ontology alignment.

Gap 3: Metadata required to support reuse decisions including mapping purpose, domain, assumptions, stakeholder involvement, design decisions, and quality results is absent or only partially captured.

Gap 4: No unified framework clarifies how metadata should be structured across project-level, mapping-level, and correspondence-level, or what level of annotation is practically feasible.

Gap 5: Recent FAIR-oriented initiatives provide valuable guidelines but do not yet define a concrete, implementable, machine-readable metadata model applicable across all mapping types.

Gap 6: Current LLM-based approaches for mapping generation across all three declarative mapping types operate without structured lifecycle metadata or guided prompt templates. Prompts are constructed in an ad hoc manner, providing no representation of mapping purpose, domain assumptions, design decisions, or quality requirements. As a result, generated outputs are semantically incomplete and difficult to interpret or reuse.

To further illustrate these limitations, Table 3.8 provides a comparative overview of the main approaches discussed in this chapter, positioning them against the proposed lifecycle-based metadata model across key dimensions.

Table 3.7 Comparative overview of existing approaches and LMMD across key dimensions

	OM2R	Project-centric	SSSOM	Alignment API	RML Approaches	FAIR-IMPACT / RDA	This thesis
Lifecycle coverage	Partial	Yes	No	No	No	Partial	Yes
Metadata granularity	Correspondence	Project to correspondence	Correspondence	Correspondence	Document/Component	Correspondence / Set	Project / Phase
Mapping types supported	Alignment only	Alignment only	Alignment only	Alignment only	Uplift only	All (conceptual)	All three
Contextual metadata	Partial	Yes	No	No	No	Partial	Yes
Reuse support	Partial	Yes	No	No	No	Partial	Yes
FAIR alignment	No	No	Partial	No	No	Yes	Yes
Machine readability	OWL-DL	OWL-2	TSV / RDF	RDF/XML	OWL / RDF	Guidelines only	OWL-DL

It is worth noting that LMMD was developed in 2022 [66], predating the FAIR-IMPACT recommendations published in 2025 [9] and the RDA FAIR Mappings Working Group endorsed in February 2025 [18]. The convergence between LMMD and these subsequent community initiatives independently is supportive of the relevance of the gaps identified in this thesis research and provides some positive indication that the research community has converged on similar requirements from different starting points.

These gaps directly motivate the design requirements for LMMD presented in Chapter 4, where each gap is explicitly addressed through a corresponding design decision.

3.5 Chapter Summary

This chapter reviewed existing work on declarative mapping lifecycles, metadata models, and FAIR-oriented initiatives. Section 3.1 traced the evolution of lifecycle models from early ontology alignment frameworks, through project-centric governance models, to FAIR-oriented initiatives. Section 3.2 examined metadata approaches across all three mapping types. Section 3.3 reviewed LLM-based mapping generation approaches. Section 3.4 identified six key gaps that collectively motivate the lifecycle-based metadata model proposed in this thesis. Chapter 4 presents the design of this model and the decisions taken to address each identified gap.

Chapter 4: Lifecycle-Based Metadata Model for Declarative Mappings (LMMD)

This chapter presents the design of the **L**ifecycle-Based **M**etadata **M**odel for **D**eclarative Mappings (**LMMD**). The design responds directly to the six gaps identify in chapter 3: the fragmentation of metadata approaches across mapping types, absence of lifecycle-structured metadata, limited support for reuse decision-making, unresolved metadata granularity, and the gap between FAIR recommendations and practical implementation. LMMD addresses these gaps through a lifecycle-based metadata model applicable across uplift mappings, interlinking specifications, and ontology alignment.

The chapter begins by clarifying the methodological approach and scope of the model, followed by outlining the key challenges observed in current declarative mapping practice. Based on these challenges, and drawing on existing approaches, intended use case scenarios, and FAIR principles, a set of requirements for lifecycle-based mapping metadata is derived.

The chapter then presents a domain analysis that clarifies how the model is intended to be used, followed by the introduction of an extended lifecycle for declarative mapping projects. Based on this foundation, the design of LMMD is described in terms of its goals, core concepts, and design features, together with an illustrative example. The chapter concludes by outlining the ontology-based implementation of the model through the Mapping Metadata Vocabulary (MMV), describing how LMMD is operationalised in the MetaSEMAP tool, and presenting the evolution of the model based on the expert feedback and iterative evaluation.

4.1 Methodology and Research Scope

The design of the Lifecycle-Based Metadata Model for Declarative Mappings (LMMD) follows a requirements-driven approach, in which design decisions are

explicitly grounded in identified challenges and intended use scenarios. The overall approach adopted in this thesis is to first analyse limitations in existing declarative mapping documentation practices and metadata approaches, and then to derive a set of design requirements that the proposed model is intended to address. These requirements are subsequently satisfied through concrete modelling decisions that are later formalised as an ontology and operationalised in a supporting tool.

This analysis highlights recurring challenges related to mapping discovery, reuse assessment, quality evaluation, and the long-term maintenance of declarative mappings, and identifies gaps in existing metadata approaches with respect to the coverage and representation of lifecycle and reuse relevant information.

In addition to the literature-based analysis presented in Chapter 3, the methodology incorporates a domain analysis that examines how declarative mappings are created, used, and revisited in practice. This analysis identifies the intended use cases for LMMD and justifies the inclusion of lifecycle phases that extend beyond mapping implementation alone.

With respect to scope, LMMD is intended to support declarative mapping projects across three common categories: uplift mappings, interlinking specifications, and ontology alignment mappings. While these mapping types differ in representation and technical detail, they share common lifecycle characteristics and documentation needs. Definitions and illustrative examples of these mapping categories are provided and discussed in detail in Chapter 2.

To provide a consistent foundation for documenting these mapping types, LMMD adopts a lifecycle-oriented perspective informed by the prior work on mapping lifecycle [38], by established software and data engineering practices [37], and by observations of how declarative mappings are developed and maintained in practice. This directly motivated the decision to organise LMMD around lifecycle phases rather than mapping artefacts alone. The six gaps identified in Section 3.4 are directly addressed through the requirements derived in Section 4.4, ensuring

that the design decisions presented in this chapter are explicitly grounded in the limitations of existing approaches.

Figure 4.1 provides an overview of the scope of LMMD, illustrating how the model provides a unified lifecycle-based documentation framework across uplift mappings, interlinking specifications, and ontology alignment, while structuring metadata according to common lifecycle phases.

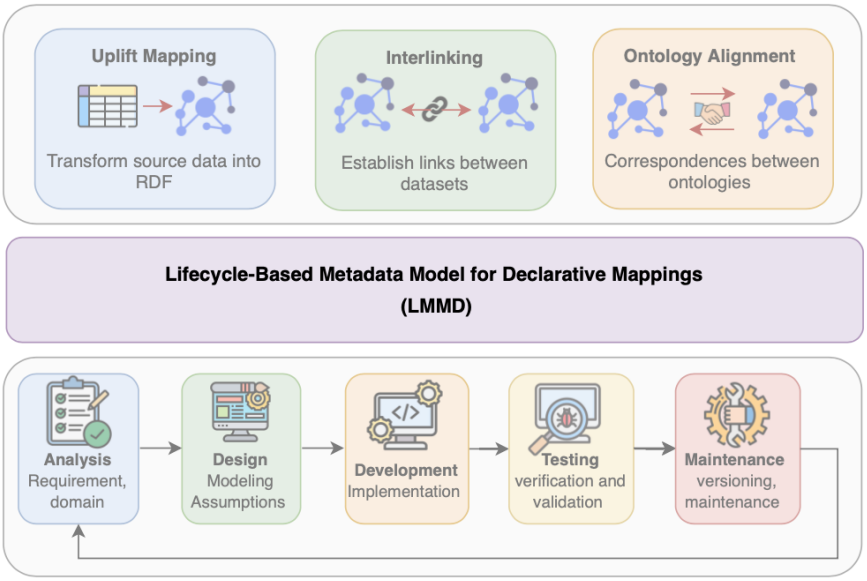


Figure 4.1 Overview of the Lifecycle-Based Metadata Model for Declarative Mappings (LMMD)

4.2 Common Challenges in Declarative Mapping Practice

Despite sustained research efforts, the state of the art analysis presented in Chapter3, shows that several challenges persist in how declarative mappings are documented, assessed, and maintained in practice. These challenges are not limited to specific mapping languages or tools, but recur across different mapping types in a wide range of application contexts. Together, they constrain mapping discovery, reuse, quality assessment, and long-term maintenance, motivating the need for a more systematic and lifecycle-aware approach to mapping metadata.

4.2.1 Type-Specific Documentation Practices

When this research was first undertaken (2022), metadata representations for declarative mappings were developed independently within uplift mapping, ontology alignment, and interlinking communities. Uplift mappings were primarily described within the R2RML/RML ecosystem [4].

As discussed in Section 3.4, existing metadata approaches across uplift mappings, ontology alignment, and interlinking evolved independently and do not share a common lifecycle-oriented abstraction. Comparable metadata such as mapping purpose, contextual assumptions, validation evidence, and maintenance status remains inconsistently represented across communities, complicating discovery, assessment, and reuse.

4.2.2 Artefact-Centric Documentation Limitations

A related challenge concerns the granularity at which metadata is represented for declarative mappings. Existing approaches tend to associate metadata either with individual correspondences or with technical generation artefacts, rather than with the broader mapping project as a lifecycle entity.

As discussed in Section 3.4 (Gap 2), existing approaches associate metadata with individual correspondences or technical artefacts rather than with the mapping project as a lifecycle entity. When mappings are revisited or assessed for reuse, users typically have access to technical metadata about how correspondences were produced, but little structured information about why the mapping was created, under which constraints, and with what validation criteria.

4.2.3 Limited Lifecycle and Validation Support

Beyond type-specific and artefact-centric documentation practices, existing approaches provide limited support for capturing metadata across the full lifecycle of declarative mappings. Documentation efforts tend to focus on implementation

and execution, with considerably less attention paid to early-stage activities such as analysis and design, or to later-stage activities such as testing, validation, and maintenance.

Validation and quality assessment are particularly underrepresented across the approaches discussed in Chapter 3. Mappings may be tested during or after the development, but evidence of such testing such as quality metric values, evaluation criteria, or testing type are often undocumented or recorded informally. This lack of documented validation makes it difficult for third parties to assess the trustworthiness or suitability of a mapping for reuse.

Long-term maintenance and evolution are also rarely addressed. Data sources, schemas, and ontologies may change over time and metadata describing versioning and updates are often absent or not documented. Governance of mappings becomes complicated and the risk of deploying or reusing outdated or incompatible mappings increases without explicit and proper documentation of these changes.

These three challenges map directly to the gaps identified in Section 3.4. The type-specific fragmentation of metadata approaches corresponds to Gap 1. The artefact-centric limitations and absence of lifecycle coverage correspond to Gaps 2 and 3. The limited lifecycle and validation support corresponds to Gaps 3 and 4. Together, they motivate the requirements derived in Section 4.4.

4.3 Domain Analysis and Intended Use Cases

To ground the derivation of requirements in practical usage contexts, this section presents a domain analysis of declarative mapping projects and identifies the intended use cases that a lifecycle-based metadata model is expected to support. Rather than focusing on specific tools or mapping languages, the analysis abstracts over common roles, activities, and information needs observed across declarative mapping practice. This domain perspective provides a concrete basis for understanding how mappings are created, used, revisited, and reused over time.

This domain analysis draws on the state of the art review, insights from prior empirical work, and practical experience gained during the development and use of mapping tools. Its purpose is not to exhaustively model the domain, but to identify those aspects of mapping practice that are most relevant for metadata design and reuse-oriented documentation.

4.3.1 Stakeholders and Roles

Declarative mapping projects typically involve multiple stakeholders with different responsibilities. These stakeholders may include mapping developers, domain experts, data engineers, ontology engineers, and downstream users who consume or reuse mapping artefacts. Additional roles may also be involved depending on the context of the mapping project.

Mapping developers focus on implementation tasks such as mapping algorithm, method, tools used, formats. Domain experts commonly contribute knowledge about the source and target data, domain assumptions, constraints and modelling decisions that influence mapping semantics. Downstream users or consumers, who may not have been involved in the original development, require sufficient contextual information about the mapping to interpret them and assess their suitability for reuse. Across all these roles, structured metadata is critical for supporting communication, interpretation, and decision-making throughout the mapping lifecycle.

4.3.2 Mapping Lifecycle Use Case Scenarios

The domain analysis identifies several recurring use case scenarios in which declarative mapping metadata plays a central role. These scenarios arise at different points in the mapping lifecycle and involve distinct information needs, yet all depend on the availability of structured, reuse-oriented metadata.

A first scenario concerns mapping **discovery for interpretation and reuse**, in which users search for existing mappings to understand how data has been transformed or linked, or to identify candidate mappings that may be reused or adapted in new contexts. In this scenario, users require access to contextual metadata describing the mapping's purpose, scope, assumptions, and development context, as such information cannot be reliably inferred from the mapping artefact alone.

A second scenario involves **mapping assessment and quality evaluation**, where users seek to determine whether a mapping is trustworthy and suitable for a particular use. This requires metadata describing validation and testing activities, quality criteria, and any risks or issues. By documenting this information in a structured and machine-readable way, reuse decisions will be based on documented evidence rather than speculation.

A third scenario is about **mapping maintenance and evolution of the mapping over time**. Existing mappings must be updated as data sources, underlying schemas, ontologies, or requirements change. Comprehensive and rich lifecycle-aware metadata model supports traceability across different versions, clarifies the rationale for changes made to the mappings, and enables informed governance decisions over time.

These use case scenarios show that metadata requirements extend beyond mapping execution alone and span multiple lifecycle phases. They motivate the need for a metadata model that supports interpretation, assessment, reuse, and long-term maintenance of declarative mappings.

Figure 4.2 illustrates these use case scenarios and the roles involved using a UML use case diagram. The diagram distinguishes between three stakeholder roles: knowledge engineers, mapping developers, and domain experts. Knowledge engineers are modelled as the primary actors interacting with mappings across the lifecycle, spanning discovery, interpretation, quality assessment, reuse, and adaptation activities. Mapping developers are represented as a specialised form of

knowledge engineer, inheriting all general use case interactions while additionally being responsible for mapping development and long-term maintenance. Domain experts are associated with mapping interpretation and assessing the quality of mapping, reflecting their role in validating semantics , assumptions, and domain-specific decisions.

Different stakeholders engage with declarative mappings at different lifecycle stages, reinforcing the need for lifecycle-aware metadata that supports diverse, role-dependent information needs.

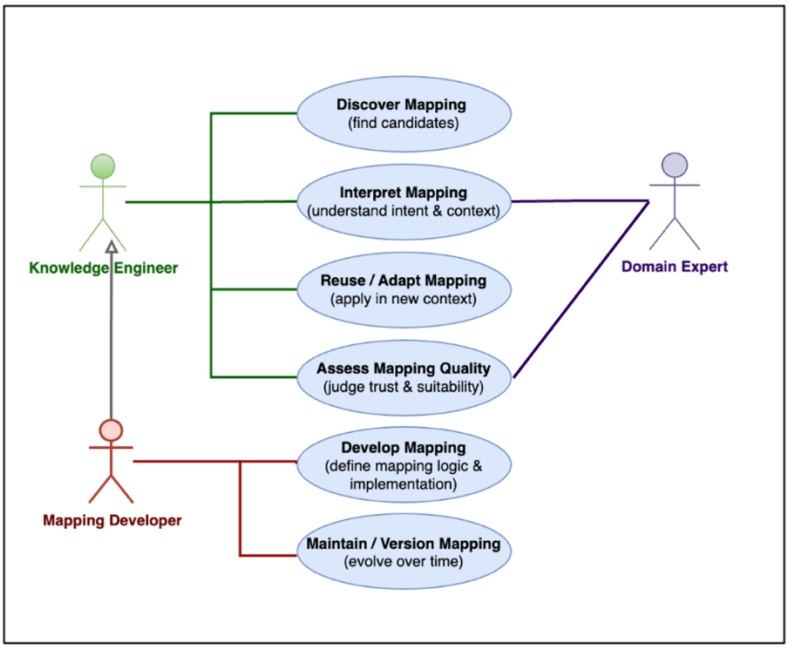


Figure 4.2 Summary of use case scenarios and the roles involved

4.3.3 Mapping Lifecycle Challenge and Design Motivation

A mapping lifecycle provides a structured view of the declarative mapping process by decomposing it into a sequence of stages. As discussed in Chapter 3, early lifecycle formulations were introduced as part of the OISIN framework [38] and later extended by Thomas [39] to support finer-grained phases and mapping reuse. More recent work, such as that of Debruyne [5], adopts a project-centric perspective on ontology mapping. While these approaches demonstrate the value of lifecycle thinking, they differ in scope and application across mapping types and

provide limited support for explicitly capturing mapping purpose, validation, and reuse-relevant context.

The state of the art analysis shows that previous approaches did not consider testing and validation as distinct lifecycle phases, despite their importance for assessing correctness and reuse suitability. To address this limitation, this work adopts and refines a lifecycle structure that is applicable across different declarative mapping types and includes testing as a dedicated phase. The proposed lifecycle serves as a conceptual foundation for the lifecycle-based metadata model introduced in this chapter. A detailed discussion and comparison of existing lifecycle models, including OISIN [38], OM2R [39], and the project-centric model [5], is provided in Chapter 3, Section 3.1.

The lifecycle is designed not only to address the limitations but also to be accessible to practitioners with diverse technical backgrounds. The lifecycle phases and the terminologies used in LMMD are **intentionally aligned** with phases from **established software engineering workflows** to reduce the learning curve and support easier adoption in practice.

The lifecycle is structured into the following phases: Analysis, Design, Mapping Development, Mapping Testing, and Mapping Maintenance. These phases capture both early decision-making activities and post-implementation validation and evolution, while remaining sufficiently generic to be applicable across different project contexts. Figure 4.3 illustrates the lifecycle adopted in this work for documenting declarative mapping projects.

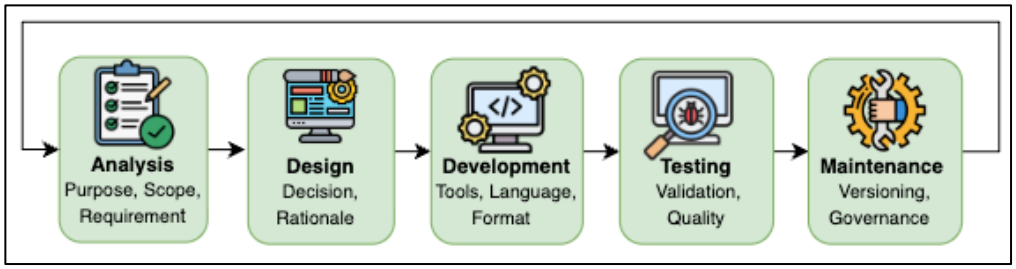


Figure 4.3 Lifecycle phases for declarative mapping projects

While the phases are presented sequentially, the lifecycle is not intended to be strictly linear. Iteration between phases is expected, particularly as mappings are revised in response to validation outcomes, evolving requirements, or changes in data sources and ontologies. This lifecycle structure provides a conceptual foundation for organising mapping metadata and informs both the domain model and the requirements derived in the subsequent sections.

4.3.4 Conceptual Domain Model and Usage Overview

Building on the identified declarative mapping use case scenarios, the domain analysis abstracts from individual tools and mapping languages to a conceptual view of how declarative mappings are situated within broader mapping projects. From this perspective, mappings are not treated as isolated executable artefacts, but as outcomes of coordinated activities that unfold over time and involve multiple stakeholders.

At a conceptual level, a **mapping project** represents the central unit of organisation. A mapping project encompasses the intent, scope, and context of a mapping effort, as well as the sequence of activities through which mapping artefacts are produced, evaluated, reused, and maintained. These activities are structured according to a lifecycle and include analysis, design, development, testing, and maintenance phases. Each activity may give rise to one or more mapping artefacts, and artefacts may be revisited or revised as the project evolves.

From a usage perspective, this project-centric abstraction supports traceability across the mapping lifecycle. It enables users to relate mapping artefacts to the activities, decisions, and evaluations that led to their creation, and to understand how mappings have evolved over time. Stakeholders interact with the mapping project at different points in the lifecycle, depending on their role and information needs, as illustrated in the use case scenarios described in Section 4.3.2.

Figure 4.4 presents a conceptual UML domain model that captures these core concepts and relationships. The model shows the central role of mapping projects, their association with lifecycle activities, the links between activities and mapping artefacts, and the involvement of different stakeholder roles. Making these relationships explicit provides a foundation for structuring lifecycle-based metadata that supports interpretation, assessment, reuse, and long-term maintenance of declarative mappings.

This abstraction informs the extended lifecycle introduced in the following section and serves as a basis for deriving concrete metadata requirements and design decisions in subsequent sections.

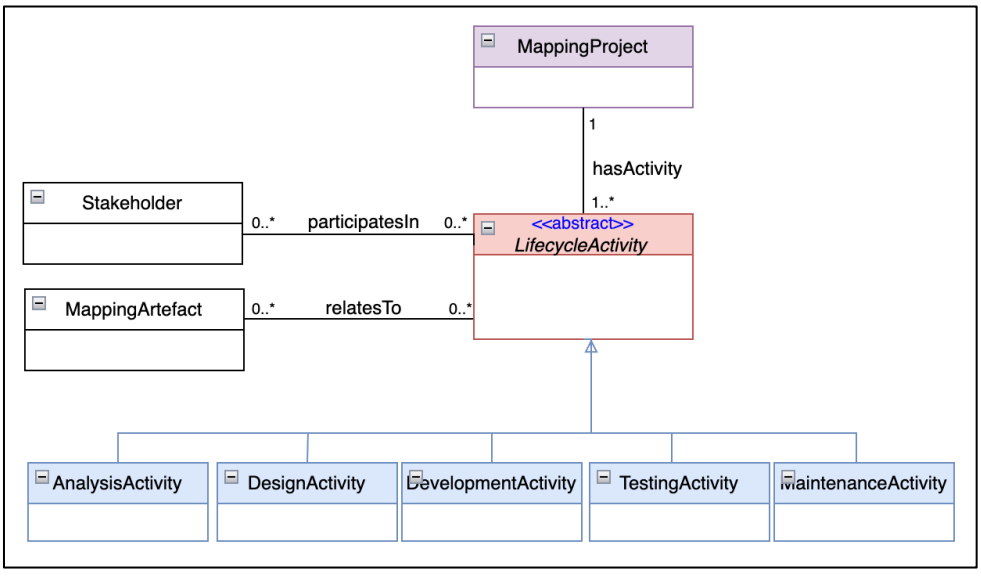


Figure 4.4 Conceptual UML domain model for lifecycle-based documentation of declarative mapping projects.

4.4 Requirements Derived From Existing Approaches, Use Case Scenarios and FAIR Principles

Based on the challenges identified in Section 4.2, the intended declarative mapping use case scenarios described in Section 4.3, and guiding considerations derived from FAIR principles, the Lifecycle-Based Metadata Model for Declarative Mappings (LMMD) is grounded in a set of design requirements. These requirements specify the capabilities that a metadata model must provide in order to support interpretation, assessment, reuse, and long-term maintenance of declarative mappings.

In this context, FAIR is interpreted as a set of guiding principles indicating that declarative mappings and their associated metadata should be findable, accessible, interoperable, and reusable. Rather than treating FAIR as a strict compliance target, these principles are used to shape expectations regarding the availability, structure, and representation of mapping metadata. The resulting requirements therefore reflect practical design considerations aligned with FAIR goals, while remaining grounded in observed practice and tool constraints.

For clarity, the requirements are organised into three categories reflecting the primary concerns emerging from the analysis: reuse and interpretation, assessment and quality, and interoperability and representation. Each requirement is mapped to one or more of the gaps identified in Section 3.4, ensuring explicit traceability between the state of the art limitations and the design decisions of LMMD. Table 4.1 summarises the resulting requirements.

Table 4.1 Requirements for the Lifecycle-Based Metadata Model for Declarative Mappings (LMMD)

Category	Requirements	Addresses Gap
Reuse and Interpretation Requirements	**R1: Lifecycle Coverage.** The metadata model must represent information across the full declarative mapping lifecycle, including analysis, design, development, testing, and maintenance. This includes both early-stage metadata such as purpose, scope, assumptions, and requirements as well as later-stage metadata covering validation, governance, and long-term maintenance.	**Gap 2**
	R2: Project-Level Representation. Metadata must describe mapping projects as first-class entities rather than focusing exclusively on individual mapping artefacts without any context about the project. This is necessary to support traceability across lifecycle activities and multiple artefacts.	**Gap 2, Gap 3**
	R3: Support for Multiple Declarative Mapping Types. The metadata model must be applicable across different declarative mapping types, including uplift mappings, interlinking specifications, and ontology alignments. Despite differences in representation and technical detail, these mapping types share a largely common lifecycle structure.	**Gap1**
Assessment and Quality-Related Requirements.	**R4: Support for Reuse Assessment and Quality Evaluation.** The metadata model must enable informed reuse decisions by supporting the assessment of mapping suitability, quality, and provenance. This requires explicit representation of validation and testing activities, quality criteria, results, and known limitations.	**Gap 3**
	R5: Traceability of Decisions and Evaluation Outcomes. Metadata should support traceability between lifecycle activities, mapping artefacts, and assessment outcomes.	**Gap 3, Gap 4**
Interoperability and Representation Requirements.	**R6: Interoperability through Standards Alignment.** Where possible, the metadata model should reuse and align with established Semantic Web and Linked Data vocabularies to promote interoperability and avoid unnecessary duplication.	**Gap 5**
	R7: Machine-Readable and Human-Understandable Representation. Metadata should be represented in a structured, machine-interpretable form while remaining accessible to human users.	**Gap 4, Gap 5**

These requirements define the scope of LMMD and provide a concrete basis for the design decisions and traceability mechanisms introduced in the subsequent sections. They operationalise FAIR principles by promoting metadata practices that improve the findability, accessibility, interoperability, and reuse of declarative mappings.

Table 4.2 summarises how the design requirements of LMMD align with the FAIR principles introduced in Section 2.6. Each FAIR principle is addressed through one or more requirements, demonstrating that LMMD is aligned with broader goals for making declarative mappings findable, accessible, interoperable, and reusable.

Table 4.2 Summary of how the design requirements of LMMD align with the FAIR principles

FAIR Principle	How LMMD addresses it	Requirements
Findable	Structured project-level metadata including purpose, domain, and identifiers supports discovery of mappings	**R1, R2**
Accessible	MetaSEMAP enables export of machine-readable metadata in open RDF formats	**R6, R7**
Interoperable	MMV aligns with established vocabularies including PROV-O, DCTERMS, and MQV	**R3, R6**
Reusable	Lifecycle context including design decisions, validation results, and stakeholder roles supports informed reuse decisions	**R1, R4, R5**

4.5 Design of the Lifecycle-Based Metadata Model

This section presents the design of LMMD by showing how the requirements derived in Section 4.4 are realised through concrete modelling decisions, providing explicit traceability between challenges, requirements, and the resulting metadata model.

4.5.1 Design Features and Requirement Traceability

Based on the requirements identified in Section 4.4, LMMD is realised through a set of design features (each uniquely identified using a Dx notation where x is a number) that together define its conceptual backbone. Each feature addresses one or more requirements and reflects explicit design decisions.

- **D1: Project-centric representation.** LMMD models a mapping project as a first-class entity that acts as the contextual anchor for associated lifecycle activities and mapping artefacts. This directly addresses project-level

representation (R2) and supports reuse and assessment by making explicit the purpose, scope, and stakeholders of mapping development (R4).

- **D2: Lifecycle-based organisation.** Metadata is structured around five lifecycle stages: analysis, design, development, testing, and maintenance. This supports full lifecycle coverage (R1) and consistent documentation across mapping types (R3).

- **D3: Separation of activities and artefacts.** LMMD distinguishes between lifecycle activities and mapping artefacts, linking them through provenance relations. This supports traceability and reuse assessment (R1, R2, R4).

- **D4: Mapping-type agnostic core.** The model provides a shared abstraction applicable across uplift mappings, interlinking specifications, and ontology alignments. Metadata is not bound to specific mapping languages or tools (R3).

- **D5: Standards reuse and alignment.** The model reuses established vocabularies such as PROV-O, Dublin Core Terms, FOAF, DQV, and MQV. New terms are introduced only when necessary. This supports interoperability and FAIR-aligned design goals (R6).

- **D6: Dual human–machine operationalisation.** LMMD supports both machine-readable representations for querying and comparison, and human-oriented annotation workflows to reduce documentation burden (R7).

Table 4.3 summarises how the LMMD design features address the design requirements.

Table 4.3 Traceability between LMMD design requirements and design features

Requirements	Design Features
R1 (Lifecycle coverage)	D2, D3, D6
R2 (Project-Level representation)	D1,D3
R3 (Multiple mapping types)	D2, D4
R4 (Reuse assessment and quality evaluation)	D1, D3, D6
R5 (Traceability of decisions and evaluation outcomes)	D3, D6
R6 (Interoperability through standards alignment)	D5
R7 (Machine-readable and human-understandable)	D6, D5

4.5.2 Lifecycle Phases and Associated Metadata

LMMD organises metadata according to five lifecycle phases: Analysis, Design, Development, Testing, and Maintenance. This organisation provides a consistent documentation structure across different declarative mapping types (R2) and encourages metadata capture at the point in the workflow where it naturally arises, rather than relying on retrospective reconstruction from the mapping artefact alone (R1, R4). Each lifecycle phase and its associated metadata are described below.

- **Analysis phase.** The analysis phase captures project intent and contextual grounding, including the purpose of the mapping, domain, requirements, stakeholder roles, and assumptions or constraints. This information is frequently absent from mapping artefacts but is essential for interpreting mappings and assessing their suitability for reuse (R1, R3, R4). In LMMD, analysis metadata also includes descriptions of input resources, as changes in inputs are a common cause of mappings becoming outdated or unsuitable for reuse.

- **Design phase.** The design phase captures modelling decisions and their rationale. These decisions strongly influence the semantics of the produced RDF or links but are rarely documented explicitly. LMMD therefore treats design rationale as first-class metadata, enabling future users to understand not only what was implemented, but why those choices were made (R1, R4).

- **Development phase.** The development phase captures implementation-related metadata such as mapping language, tools used, mapping methods or algorithms. While such metadata is often partially available in existing tools, LMMD integrates it into a broader lifecycle context to enable the interpretation of implementation in relation to earlier requirements and later validation results (R1, R2).

- **Testing phase.** The testing phase captures validation and evaluation activities, including test types, quality metrics, results, and identified issues. Explicit testing metadata addresses a recurring limitation identified in the state of the art: mappings are often deployed without evidence of validation. Capturing this information supports quality assessment, comparison between mapping alternatives, and reuse decisions that depend on trust (R4).

- **Maintenance phase.** The maintenance phase captures metadata related to evolution over time, including versioning, publication information, and links to released artefacts. Since declarative mappings evolve as data sources and ontologies change, maintenance metadata is essential for supporting long-term governance and reuse (R1, R3, R4).

Table 4.4 presents the initial set of metadata fields proposed in the first published version of the model [66], which served as the conceptual starting point for LMMD. The model was subsequently refined and extended through iterative evaluation, as described in Section 4.8. The full specification of the evolved model, including ontology property mappings, is presented in Section 4.6.3.

Table 4.4 Metadata fields captured across LMMD lifecycle phases,

Mapping Lifecycle Phases	Proposed Metadata Fields
Analysis	Metadata about stakeholders (URI, Name, Background, Role, organization),
	Metadata about the purpose of the mapping (Requirements, Type of mapping, Mapping domain, Domain assumptions, technical requirement, Risks, or issues),
	and Metadata to describe inputs that will be mapped (URI, Name, Source, Type, Creator, Format).
Design	(Final design decisions, Design decision justification, Quality metrics to consider during the development)
Development	(URI, Name, Start/End date of the development, Tools, Mapping method, Mapping algorithm, Format)
Testing	(URI, Name, Testing type, Date/Time, Testing result)
Maintenance	(Publisher name, Publisher source, Version number, Version date/time).

4.5.3 Competency Questions as Design Validation Instrument

Competency questions (CQs) serve as a design validation mechanism in LMMD, ensuring the model addresses the information needs associated with declarative mapping interpretation, assessment, reuse, and long-term maintenance. In LMMD, competency questions are not treated as the primary source of requirements; rather, they are used to validate the requirements derived from the state-of-the-art analysis and the intended usage scenarios described in Section 4.3. While the requirements define what LMMD must support, the competency questions provide a concrete means of assessing whether the emerging model structure enables those requirements to be addressed in practice.

The competency questions were authored by the thesis researcher as an internal validation instrument, formulated by synthesising (i) gaps and limitations identified in the literature reviewed in Chapters 2 and 3, and (ii) recurring questions that practitioners typically pose when encountering an existing mapping artefact.

During the design process, competency questions were used iteratively to guide modelling decisions and assess coverage. Concepts and relationships were introduced, refined, or excluded based on their contribution to answering one or more competency questions. This iterative use of CQs ensured that LMMD

evolved in a controlled manner, with each modelling decision justified by its relevance to explicit information needs rather than by ad hoc or purely conceptual considerations.

Table 4.5 presents a **subset** of the competency questions used to guide the design of LMMD. These questions show the range of reuse-oriented queries the model is designed to support across different lifecycle phases.

Table 4.5 Representative competency questions used to guide the design of LMMD

CQ	Competency Question
CQ1	What is the purpose and scope of the mapping project?
CQ2	Which mapping type is being documented?
CQ3	What assumptions or requirements does the mapping rely on?
CQ4	Which lifecycle activities are associated with the mapping project?
CQ5	What key design decisions were made and why?
CQ6	Which mapping artefacts were produced and how are they described?
CQ7	Which mapping language and format were used?
CQ8	Which tools or methods were used during development?
CQ9	How was the mapping tested and what were the results?
CQ10	How is the mapping maintained and versioned over time?

The complete set of competency questions, together with their detailed mappings to MMV classes and properties (which are introduced next in Section 4.6) and example SPARQL query templates used during validation, is provided in **Appendix B.**

4.6 Ontology Implementation and Tool Operationalisation

This section presents how the Lifecycle-Based Metadata Model for Declarative Mappings (LMMD) is implemented as a formal ontology through the Mapping Metadata Vocabulary (MMV). While the previous sections focused on requirements, lifecycle motivation, and conceptual structure, the purpose of this section is to explain how these design decisions are expressed in machine-

interpretable vocabulary that enables lifecycle-based documentation, assessment, and reuse of declarative mapping projects in practice.

MMV is implemented as an OWL 2 DL ontology and published at https://sarahalzahranitcd.github.io/MMV-Ontology2/ (Version 0.1, released 2026-02-01). The vocabulary is versioned, openly licensed, and accompanied by human-readable documentation generated using WIDOCO. Rather than introducing a new provenance or quality framework, MMV complements existing vocabularies by focusing specifically on mapping-centric lifecycle metadata that is not adequately covered elsewhere. The ontology was introduced and presented at the IEEE ICSC 2026 [83].

4.6.1 Ontology Engineering Approach

The development of the Mapping Metadata Vocabulary (MMV) followed a requirements-driven ontology engineering approach, informed by established methodologies in the Semantic Web community rather than adhering rigidly to a single prescriptive framework. In particular, the design was influenced by the Linked Open Terms (LOT) methodology [67], which structures ontology development into requirements specification, implementation, publication, and evaluation stages. Reuse-oriented design decisions were guided by the NeOn framework [68], which recommends aligning with existing vocabularies and introducing new terms only where existing standards are insufficient.

These methodologies were applied selectively, reflecting the exploratory and design-oriented nature of this research. MMV was developed iteratively, with continuous refinements driven by requirement validation, competency question coverage, and feedback obtained during implementation of the MetaSEMAP tool (see section 4.7) and through the community survey (Study 1, Section 5.3), and expert feedback (Study 3, Section 5.5). This approach ensured that methodological guidance supported the design of a practical and reuse-oriented ontology.

The ontology development process proceeded as follows:

1. The scope, aims, and requirements of the ontology were established based on the state-of-the-art analysis in Chapter 3 and the usage scenarios in Section 4.3.
2. Competency questions were formulated to validate coverage during ontology design, as described in Section 4.5.3.
3. Existing vocabularies were identified and analysed for reuse, including PROV-O [11], Dublin Core Terms [10], FOAF, DQV, and MQV [9].
4. MMV classes and properties were identified and modelled iteratively until each competency question was satisfied.
5. The ontology was implemented in OWL 2 DL using Protégé. OWL 2 DL was selected for its broad tooling support and syntactic expressivity rather than for runtime inference. MMV defines disjointness axioms between the five lifecycle activity classes, domain and range restrictions on all properties, and inverse property pairs to support bidirectional traversal. These axioms support consistency checking during ontology development rather than producing inferred triples at query time. A reasoner was applied during development to verify that all classes were satisfiable and that no disjointness violations existed.
6. The ontology was refined iteratively based on feedback from MetaSEMAP implementation, the community survey (Study 1, Section 5.3), and expert evaluation (Study 3, Section 5.5).
7. MMV was evaluated using FOOPS! and OOPS!, as described in Section 5.5.
8. Ontology documentation was generated using WIDOCO and published under a CC BY 4.0 licence.

4.6.2 Mapping the Conceptual Model to OWL (MMV)

The MMV ontology is structured around three interrelated modelling layers: **mapping projects**, **lifecycle activities**, and **mapping artefacts**. This structure

directly reflects the conceptual design presented in Section 4.3 and operationalises the lifecycle-based organisation of metadata.

- At the core of the ontology is the class ***mmv:MappingProject***, which represents a declarative mapping as a first-class entity. A mapping project acts as the contextual anchor for all associated metadata, linking together lifecycle activities and the artefacts they produce, evaluate, or maintain. This project-centric modelling decision ensures that metadata describing purpose, scope, stakeholders, validation, and evolution is explicitly represented which directly addresses Gap 2 and R2, by ensuring that contextual information is represented at the project level rather than inferred from executable artefacts.

- Lifecycle activities are modelled as subclasses of ***prov:Activity*****,** inheriting its provenance semantics while specialising it for mapping-specific lifecycle phases. This design choice ensures that MMV metadata remains compatible with any provenance-aware system that understands PROV-O. Five activity classes are explicitly typed according to the lifecycle phases defined in LMMD: ***mmv:AnalysisActivity, mmv:DesignActivity, mmv:DevelopmentActivity, mmv:TestingActivity,*** and ***mmv:MaintenanceActivity***. Each activity class captures phase-specific metadata. Analysis activities record purpose, assumptions, requirements, and risks. Design activities capture decisions and justifications. Development activities document tools, formats, and algorithms. Testing activities record validation and quality assessment. Maintenance activities capture publication and versioning. Mapping projects are linked to these activities through specific object properties (e.g., *mmv:hasAnalysis*, *mmv:hasDesign*), enabling explicit traversal of the mapping lifecycle. The decision to type activities explicitly according to lifecycle phase rather than using a single generic *prov:Activity* class was motivated by the need to support phase-specific querying. A single activity class would not allow a system or practitioners to distinguish whether metadata was captured during the analysis phase or the testing phase, which is essential for supporting the use cases described in Section 4.3.2.

- Concrete mapping specifications are represented using the class *mmv:MappingArtifact*. Mapping artefacts include outputs such as RML mappings, ontology alignments, and interlinking specifications. The *mmv:hasMappingArtifact* property links the project to its documented artefact. Since the model enforces a one-to-one constraint between project and artefact, traceability to development and testing activities is established through the shared project context rather than direct artefact-to-activity links. Artefact-level metadata includes version, identifier, and description. Mapping type, format, and method are captured in the development activity. The separation between *mmv:MappingArtifact* and lifecycle activities directly implements Design Feature D3 from Section 4.5.1, ensuring traceability between what was produced and the activities that led to its creation. This allows a consuming system to answer competency questions such as CQ6 (which artefacts were produced) and CQ9 (how was the mapping tested).

Figure 4.5 presents a high-level diagram of the MMV ontology, illustrating the core relationships between mapping projects, lifecycle activities, and mapping artefacts. The diagram abstracts from lower-level quality and provenance constructs to emphasise the lifecycle-oriented structure that underpins LMMD.

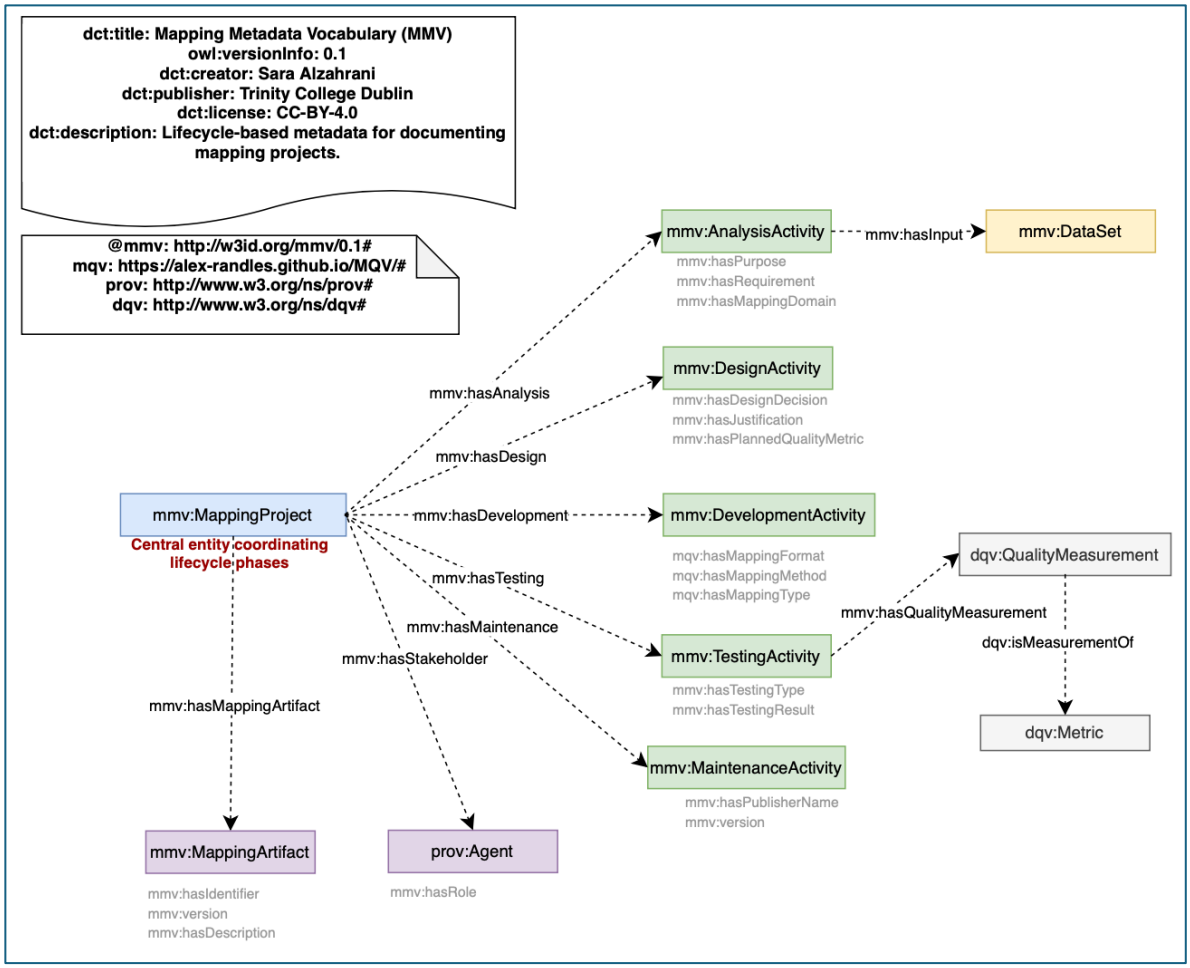


Figure 4.5 Overview of the MMV ontology

The ontology diagram shown in Figure 4.5 was created using Chowlk [69], a framework that supports the visual design of OWL ontologies using a UML-based notation and enables automatic conversion of diagrams to OWL serialisations.

4.6.3 Vocabulary Reuse and New MMV Terms

MMV follows a reuse-oriented ontology engineering approach. Existing Semantic Web and Linked Data vocabularies are reused wherever they are suitable, and new terms are only introduced when no existing standard covers the concept. For provenance, MMV aligns with PROV-O. This means MMV metadata can work with any provenance-aware system without redefining basic provenance concepts.

Descriptive metadata such as identifiers and titles reuses Dublin Core Terms. Stakeholder information uses FOAF.

For quality-related metadata, MMV reuses the Data Quality Vocabulary (DQV) and the Mapping Quality Vocabulary (MQV). DQV provides properties for describing metrics, measurements, and assessments, which MMV uses during the testing phase. MQV is reused to describe mapping-specific properties such as mapping type, mapping format, and mapping method [9].

Two design decisions are worth noting: First, ***mmv:hasJustification*** captures project-level design rationale, specifically the reasons behind modelling decisions made during the design phase. This is different from SSSOM's mapping justification field, which operates at correspondence level and describes how an individual mapping correspondence was established. The two fields address different granularities and are complementary rather than overlapping. Second, the decision to introduce ***mmv:DataSet*** instead of reusing *dcat:Dataset*. DCAT is designed for publishing catalogued dataset with persistent identifiers. That is not what is needed here. *mmv:DataSet serves a different purpose,* it records lightweight documentation about the inputs used during a specific mapping project. The dataset may not be published or catalogued at all. Requiring practitioners to produce full DCAT catalogue for every input would make the annotation impractical.

Where existing vocabularies do not provide sufficient support for lifecycle-specific concepts, MMV introduces a set of domain-specific classes and properties. The following list describes these new terms, grouped by lifecycle phase. The complete specification of all new terms is provided in Appendix C.

Project level: At project level, *mmv:MappingProject* acts as the central documentation entity, linking all lifecycle phases and the mapping artefact through object properties. Inverse properties support bidirectional traversal which allows queries to start either from the project or from individual lifecycle activities. The following properties are introduced for this level:

- Object properties to link lifecycle phases to *mmv:MappingProject*: *mmv:hasAnalysis*, mmv:hasDesign, *mmv:hasDevelopment, mmv:hasTesting, and mmv:hasMaintenance.*
- Inverse properties : *mmv:isAnalysisOf, mmv:isDesignOf, mmv:isDevelopmentOf, mmv:isTestingOf, mmv:isMaintenanceOf, and mmv:isMappingArtifactOf,* support bidirectional traversal.
- *mmv:MappingArtifact* represents a concrete mapping specification, linked to the project through *mmv:hasMappingArtifact.* Artefact-level metadata includes *mmv:hasIdentifier* for persistent identification, *mmv:version, mmv:hasDescription,* and *mmv:hasVersionDateTime.*

Analysis phase: The *mmv:AnalysisActivity* class represents the initial project phase, extending prov:Activity to capture analysis-specific metadata. The following properties are introduced for this phase:

- *mmv:hasPurpose*, captures the goal or objective of the mapping project
- *mmv:hasRequirement,* captures functional or data requirements the mapping must fulfil
- *mmv:hasMappingDomain,* captures the domain of the mapping.
- *mmv:hasDomainAssumption,* captures assumptions about the source or target data domain , such as ontological conventions or data modelling expectations.
- *mmv:hasAssumption,* captures general assumptions made during analysis about data stability, execution context, or workflow constraints.
- *mmv:hasTechnicalRequirement,* captures technical constraints or system expectations
- *mmv:hasRiskOrIssue,* captures known risks or potential issues identified during analysis
- *mmv:hasStakeholder,* links the project to a stakeholder agent, with *mmv:hasRole* and *mmv:hasBackground* describing their role and expertise. Stakeholder URI and organisation are captured through *foaf:homepage* and *foaf:organization* on the *prov:Agent* instance, supporting persistent stakeholder identification.
- Input datasets are represented as instances of mmv:DataSet, linked to the analysis activity through *mmv:hasInput.* Each dataset is described using: *mmv:hasInputName, mmv:hasInputSource, mmv:hasInputType, mmv:hasInputCreator, mmv:hasInputFormat,* and *mmv:hasInputDescription.*

Design phase: The *mmv:DesignActivity* class captures modelling and decision-making activities. The following properties are introduced:

- *mmv:hasDesignDecision*, captures key modelling decisions made during design
- *mmv:hasJustification*, captures the rationale behind each design decision

- *mmv:hasPlannedQualityMetric*, links planned quality metrics to the design phase, supporting traceability between design intent and testing outcomes

Development phase: The *mmv:DevelopmentActivity* class captures details that are related to the actual implementation of mapping artefact. The following properties are introduced:

- *mmv:hasTool*, captures the tools used during development
- *mmv:hasMappingAlgorithm,* captures the algorithm used
- Mapping type, format, and method are captured through the reused MQV properties *mqv:hasMappingType, mqv:hasMappingFormat,* and *mqv:hasMappingMethod.*
- Development start and end times are captured through *prov:startedAtTime* and *prov:endedAtTime*, inherited from *prov:Activity*.

Testing phase: The *mmv:TestingActivity* class captures validation and quality assessment activities. The following properties are introduced:

- *mmv:hasTestingType,* captures the category of testing conducted
- *mmv:hasTestingResult,* captures the outcome of the testing activity
- Quality assessment is modelled through *mmv:QualityAssessment,* linked to DQV quality measurements through *mmv:hasQualityMeasurement.*

Maintenance phase: The *mmv:MaintenanceActivity* class captures post-deployment governance metadata. The following properties are introduced:

- *mmv:hasPublisherName,* captures the name of the person or organisation that published the mapping
- *mmv:hasPublisherSource,* captures where the published mapping can be traced to

Compared to the initial conceptual model proposed in [66] (Table 4.4), **MMV v0.1** introduces three refinements. Conceptual field groupings were decomposed into individually named, formally defined properties. Each field was mapped to an existing vocabulary where possible, with new *mmv* terms introduced only where existing standards were insufficient. Two new classes were introduced based on feedback from the community survey described in (Study 1, Section 5.3): *mmv:DataSet*, which formally represents input resources as named entities, and *mmv:QualityAssessment,* which captures quality assessment activities associated with the testing phase. This reuse and extension strategy ensures that MMV remains interoperable and aligned with standards.

4.6.4 Illustrative Example

This section presents a simplified example based on the Mapping Metadata Vocabulary (MMV) to illustrate how the model can be used in practice to annotate mapping artefacts. The example describes a declarative mapping project whose objective is to uplift a tabular dataset of geospatial administrative regions into RDF for integration into a knowledge graph. The project is represented as an instance of *mmv:MappingProject*, which acts as the contextual anchor for all associated lifecycle metadata. A partial instantiation of MMV is shown in Listing 4.1.

Listing 4.1 Example instantiation of a lifecycle-based mapping project using MMV

```
mmv:Proj1 a mmv:MappingProject ;
    dcterms:title "Administrative Region Uplift Mapping Project" ;
    mmv:hasMappingArtifact mmv:Map_V1 ;
    mmv:hasAnalysis mmv:Proj1_Analysis ;
    mmv:hasDesign mmv:Proj1_Design ;
    mmv:hasDevelopment mmv:Proj1_Development ;
    mmv:hasTesting mmv:Proj1_Testing ;
    mmv:hasMaintenance mmv:Proj1_Maintenance .

mmv:Proj1_Analysis a mmv:AnalysisActivity ;
    mmv:hasPurpose "Transform tabular administrative data into RDF for
    knowledge graph integration" ;
    mmv:hasRequirement "Support reuse across multiple regional datasets" ;
    mmv:hasAssumption "Input data follows a stable CSV schema" ;
    mmv:hasRiskOrIssue "Schema changes may invalidate column mappings".

mmv:Proj1_Design a mmv:DesignActivity ;
    mmv:hasDesignDecision "Use hierarchical URIs for administrative units" ;
    mmv:hasJustification "Supports interoperability with existing spatial
    vocabularies" .

mmv:Proj1_Development a mmv:DevelopmentActivity ;
    prov:startedAtTime "2024-03-01T00:00:00"^^xsd:dateTime ;
    prov:endedAtTime "2024-06-30T00:00:00"^^xsd:dateTime ;
    mmv:hasMappingAlgorithm "Rule-based column-to-entity transformation" ;
    mqv:hasMappingType "Uplift Mapping" ;
    mqv:hasMappingMethod "Automatic" ;
    mqv:hasMappingFormat "Turtle" ;
    mmv:hasTool "RMLMapper" .

mmv:Map_V1 a mmv:MappingArtifact ;
    mmv:version "1.0" ;
    mmv:hasDescription "RML mapping for generating RDF from regional CSV data".

mmv:Proj1_Testing a mmv:TestingActivity ;
    mmv:hasTestingType "Structural validation" ;
    mmv:hasTestingResult "All required entities generated successfully".

mmv:Proj1_Maintenance a mmv:MaintenanceActivity ;
    mmv:hasPublisherName "Trinity College Dublin" ;
    mmv:hasPublisherSource "https://SEMAPrepositry.come "^^xsd:anyURI.
```

The metadata in Listing 4.1 is structured around the lifecycle phases defined in LMMD. **Lines 1-8** define the mapping project itself and link it to its analysis, design, development, testing, and maintenance activities using the object properties mmv:hasAnalysis, mmv:hasDesign, mmv:hasDevelopment, mmv:hasTesting, and mmv:hasMaintenance. This project-centric structure provides a common point through which all lifecycle-related metadata can be accessed and queried.

Lines 10-14 represent the analysis phase through an instance of mmv:AnalysisActivity. This block captures early-stage information that is typically not available in mapping artefacts alone such as the purpose of the mapping, requirements, domain assumptions, and identified risks. This supports later interpretation of the mapping and provides essential context for reuse decisions.

The design phase is represented in **lines 16-18** by a mmv:DesignActivity. This activity records key modelling decisions together with their justifications. Recording design rationale enables future users to understand why particular semantic choices were made without inferring that from the structure of the generated RDF.

Implementation-related metadata is documented in **lines 20-27** through a mmv:DevelopmentActivity. This phase records how the mapping was implemented in practice, including the mapping algorithm, format, and tools used. The development activity provides a bridge between conceptual design decisions and the actual mapping artefact produced.

The resulting declarative mapping is represented in **lines 29-31** as a mmv:MappingArtifact. The artefact is described using its version number and a brief description. This supports traceability between lifecycle activities and the concrete artefact that is executed or reused.

Testing and long-term governance are captured through the remaining lifecycle phases. **Lines 33-35** describe a mmv:TestingActivity, which records the type of testing performed and the outcome of the testing. **Lines 37-39** represent a mmv:MaintenanceActivity, capturing publication-related metadata that supports versioning and long-term management of the mapping project.

This example demonstrates how MMV integrates project-level context, lifecycle activities, and mapping artefacts into a single coherent structure. Complete formal metadata instantiation generated according to MMV for a project can be generated through the MetaSEMAP tool (see next).

4.7 MetaSEMAP Design and Implementation

An early prototype, a tool called MetaMap, was introduced at IEEE ICSC 2024 [70]. That prototype focused on supporting lifecycle documentation for uplift mappings. MetaSEMAP extends this initial prototype by supporting multiple mapping types and structured lifecycle-based metadata annotation aligned with the LMMD model. This section describes MetaSEMAP[5], its purpose, workflow and key functionalities. The version of MetaSEMAP described in this section is the initial implementation used across the evaluation studies. At this stage, the tool used a lightweight custom vocabulary based on the metag: prefix rather than the full MMV ontology, reflecting the practical need to deploy and evaluate the tool before ontology formalisation was complete.

4.7.1 Purpose and role of MetaSEMAP

LMMD is operationalised through MetaSEMAP, a guided web-based annotation tool that supports lifecycle-based metadata documentation for uplift mappings, ontology alignment, and interlinking. MetaSEMAP serves as a proof of feasibility

[5] https://metasemap-1-sarah-alzahrani.replit.app/

for the proposed design of the model by demonstrating that the model can be applied in real mapping workflows without requiring users to interact directly with ontology languages or RDF serialisations.

The tool embeds LMMD model structure within its annotation workflow. Each metadata field in the user interface corresponds to a concept defined in the model, with the initial implementation using the *metag:* vocabulary and subsequent versions aligned with MMV. This allows users to produce ontology-compliant metadata without prior knowledge of OWL or RDF. This directly supports R7 (Section 4.4). The tool offers two entry points: annotating a new mapping project from scratch, or reusing metadata from an existing annotation.

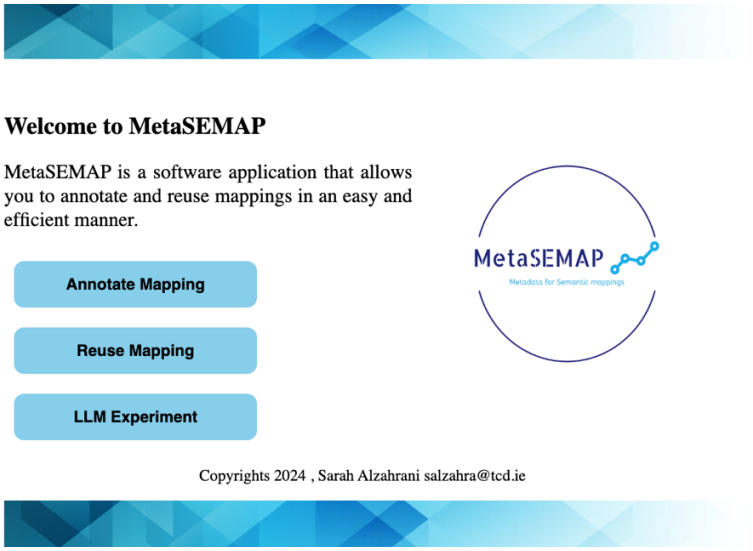


Figure 4.6 MetaSEMAP main interface

The interface shown in Figure 4.6 supports three modes corresponding to the evaluation studies conducted in this thesis: metadata annotation for Study 2, mapping reuse for Study 4, and LLM experiment described in Chapter 6. Each mode is described in the corresponding study section in Chapter 5.

4.7.2 Core annotation workflow

MetaSEMAP adopts a simple guided, web-based annotation workflow aligned with the lifecycle phases defined in LMMD. The annotation process consists of four key steps. Users begin by uploading an RDF mapping file. They then complete structured metadata fields organised according to the analysis, design, development, testing, and maintenance phases. Figure 4.7 illustrates the main steps of the annotation workflow.

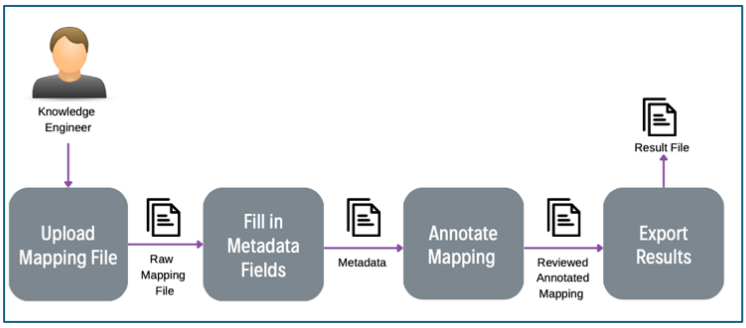


Figure 4.7 MetaSEMAP annotation workflow for lifecycle-based mapping metadata

The form is organised into sections that correspond to each lifecycle phase. Some fields are optional and scenario-specific, while others are required to ensure metadata completeness and consistency. Figure 4.8 shows the metadata annotation interface.

Contextual help is provided through tooltips attached to each metadata section. Hovering over a question mark icon reveals a short explanation of the purpose and expected input for the associated metadata section. This feature was especially useful for users and helped ensure more consistent and accurate annotations. Figure 4.9 illustrates an example tooltip.

After completing the metadata form, users are presented with a review screen summarising all the fields they have entered, as shown in Figure 4.10. Once confirmed, the tool generates machine-readable metadata in either RDF-star or Named Graph format for preview and download, as shown in Figure 4.11. The tool

was evaluated using real-world mapping scenarios based on the Virtual Record Treasury of Ireland (VRTI) knowledge graph, covering all three supported mapping types. These scenarios are described in detail in Study 2 (Section 5.4).

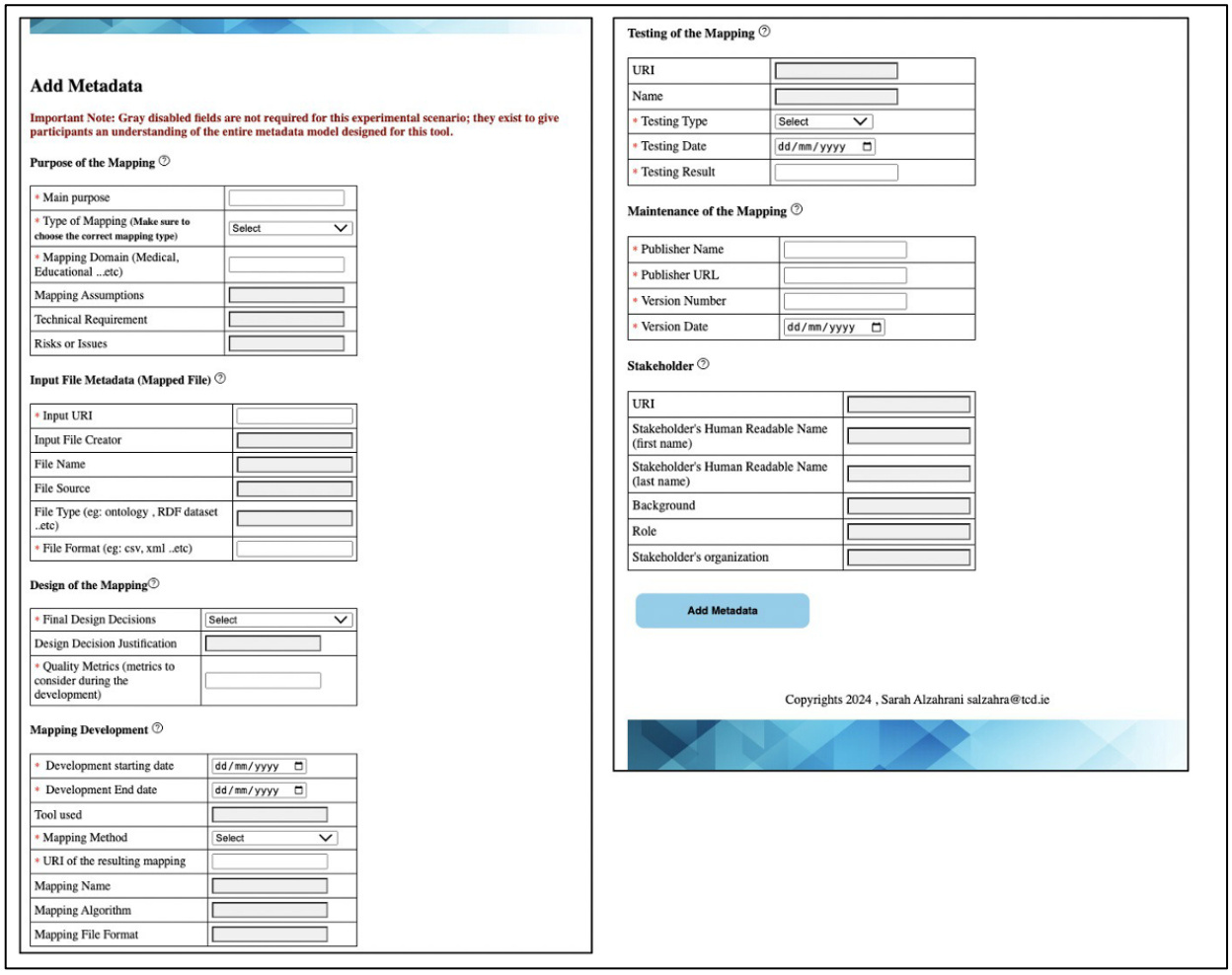


Figure 4.8 MetaSEMAP metadata annotation interface

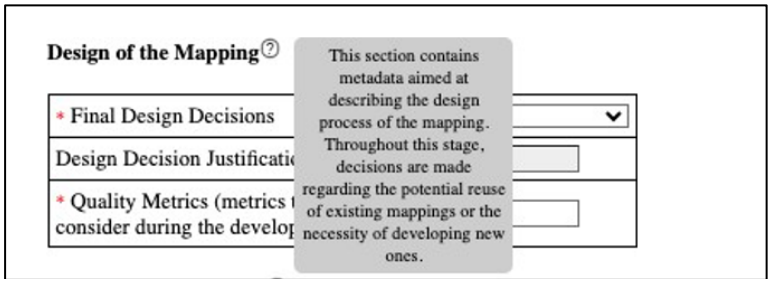


Figure 4.9 Tooltip example: metadata field explanations

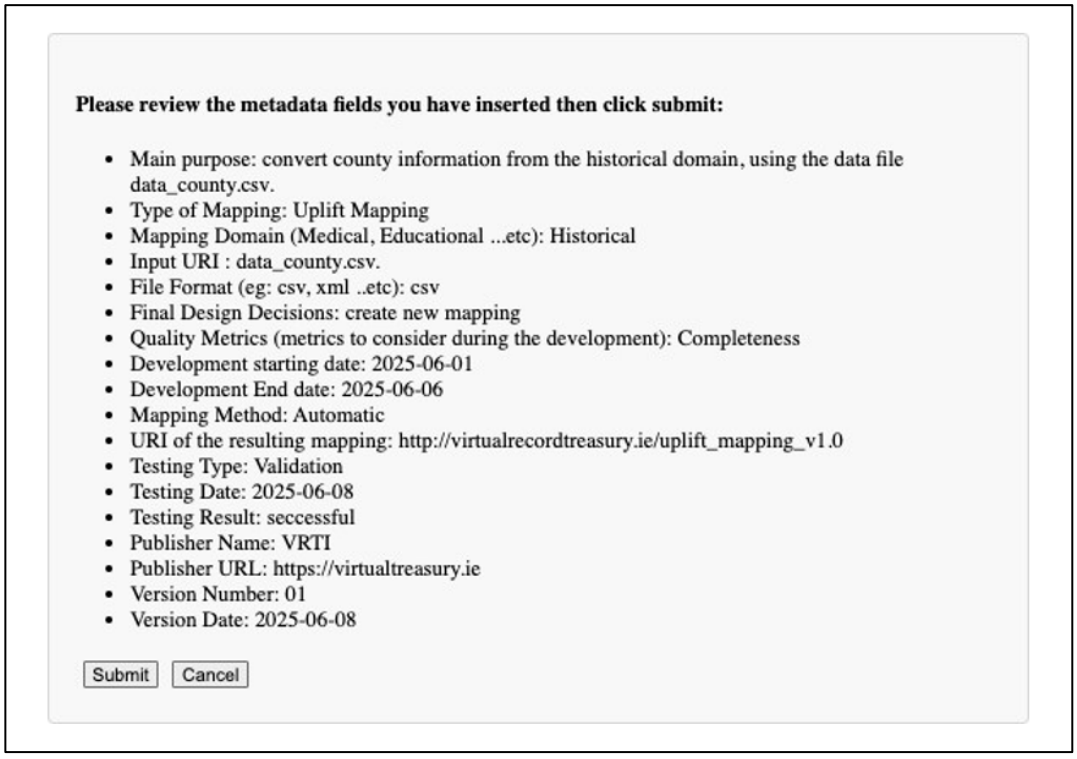


Figure 4.10 MetaSEMAP review screen displaying a summary of the entered metadata fields

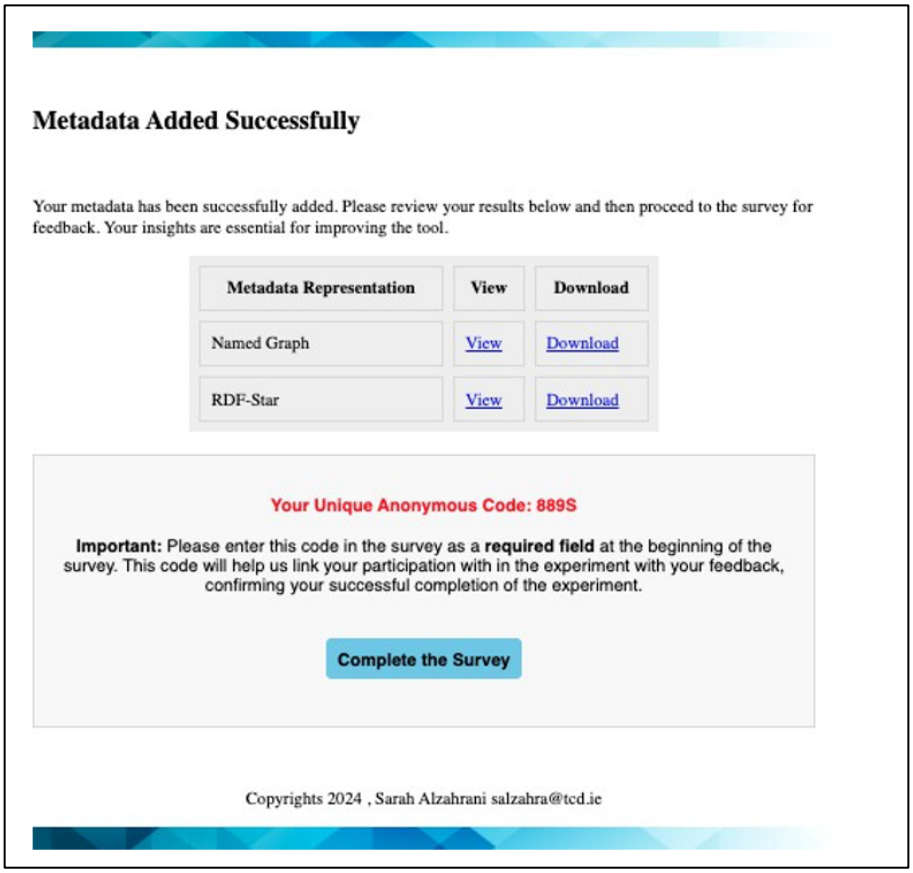


Figure 4.11 MetaSEMAP post-annotation output screen

4.7.3 Supported metadata representations

Metadata captured through MetaSEMAP can be exported in machine-readable form using two RDF-based representations: RDF-star and Named Graphs. The choice of these two representations was informed by a benchmarking studies[46] [41]. Based on these studies, RDF-star and Named Graphs offered the best balance of expressivity, query performance and tooling compatibility for mapping lifecycle metadata. Both approaches share a common conceptual model but differ in how metadata statements are encoded and stored.

To illustrate how MetaSEMAP generates output for each supported mapping type, this section presents abbreviated examples of the RDF-star and Named Graph annotations produced by the tool. Each example uses the VRTI scenarios described in Study 2 (Section 5.4). Full examples are available in the MetaSEMAP project repository. The annotation outputs shown in Listings 4.2-4.7 were generated by the

initial version of MetaSEMAP using the metag: prefix-based vocabulary rather than the full MMV ontology. This reflects the implementation state at the time of the evaluation studies. The metag: terms map conceptually to the corresponding MMV properties, for example, metag:purpose corresponds to mmv:hasPurpose, and metag:mappingType corresponds to mqv:hasMappingType. Following the evaluation studies, the tool was updated to align with MMV v0.2. Based on the evaluation of Study 2 further refinments applied to the tool. In the new version, no mapping need to be upload and the tool can be used to creating a metadata for any mapping type. An updated version available at MetaSEMAP[6] website along with all the experiment desingned in this thesis.

Uplift mapping

Listing 4.2 shows the RDF-star annotation for an uplift mapping. The metadata is attached to the triple that declares the primary TriplesMap entity using the <<...>> notation, embedding the annotation inline within the same file as the mapping. Listing 4.3 shows the Named Graph output for the same uplift mapping. The metadata is written as a separate Turtle file, with *metag:subject* as a placeholder for the mapping URI. This approach was an initial implementation choice made to enable rapid deployment and evaluation. MetaSEMAP v2 addresses this by producing conformant TriG output with a proper named graph URI, which formally separates the metadata graph from the default graph and enables graph-level querying.

[6] https://metasemap-1-sarah-alzahrani.replit.app/

Listing 4.2 RDF-star annotation for an uplift mapping using initial metag: vocabulary

```
@prefix rr:    <http://www.w3.org/ns/r2rml#> .
@prefix rml:   <http://semweb.mmlab.be/ns/rml#> .
@prefix metag: <http://example.com/metag/> .
@prefix foaf:  <http://xmlns.com/foaf/0.1/> .
@prefix xsd:   <http://www.w3.org/2001/XMLSchema#> .

<http://example.com/ns#COUNTY> a rr:TriplesMap ;
    rml:logicalSource [ rml:source "data_county.csv" ;
                        rml:referenceFormulation ql:CSV ] ;
    rr:subjectMap [ rr:template
        "https://kg.virtualtreasury.ie/place/county/{NAME_TAG}/{VRTI_ID}" ] .

<< <http://example.com/ns#COUNTY>
   rdf:type rr:TriplesMap >>
    metag:purpose      "Convert county information from historical data into
RDF
                        triples to enrich the VRTI Knowledge Graph." ;
    metag:mappingType  "Uplift Mapping" ;
    metag:mappingDomain "Historical" ;
    metag:startDate    "2024-01-01"^^xsd:date ;
    metag:endDate      "2024-02-28"^^xsd:date ;
    metag:mappingMethod "Automatic" ;
    metag:mappingURI   <http://virtualrecordtreasury.ie/uplift_mapping_v1.0> ;
    metag:testingType  "Validation" ;
    metag:testingResult "Successful" ;
    metag:publisherName "VRTI team" ;
    metag:versionNumber "1.0" ;
    metag:versionDateTime "2024-02-28"^^xsd:date .
```

Listing 4.3 Named Graph annotation for an uplift mapping

```
@prefix metag: <http://example.com/metag/> .
@prefix foaf:  <http://xmlns.com/foaf/0.1/> .
@prefix xsd:   <http://www.w3.org/2001/XMLSchema#> .

metag:subject
    metag:purpose "Convert county information from historical data into RDF
                   triples to enrich the VRTI Knowledge Graph." ;
    metag:mappingType   "Uplift Mapping" ;
    metag:mappingDomain "Historical" ;
    metag:startDate     "2024-01-01"^^xsd:date ;
    metag:endDate       "2024-02-28"^^xsd:date ;
    metag:mappingMethod "Automatic" ;
    metag:mappingURI    <http://virtualrecordtreasury.ie/uplift_mapping_v1.0> ;
    metag:testingType   "Validation" ;
    metag:testingResult "Successful" ;
    metag:publisherName "VRTI team" ;
    metag:versionNumber "1.0" ;
    metag:versionDateTime "2024-02-28"^^xsd:date .
```

Ontology alignment

Listing 4.4 shows the RDF-star annotation for an ontology alignment file. The annotation subject is the *align:Alignment* instance, which represents the alignment artefact. Listing 4.5 shows the Named Graph equivalent for the ontology alignment.

Listing 4.4 RDF-star annotation for an ontology alignment

```
@prefix align: <http://knowledgeweb.semanticweb.org/heterogeneity/alignment#> .
@prefix foaf:  <http://xmlns.com/foaf/0.1/> .
@prefix metag: <http://example.com/metag/> .
@prefix xsd:   <http://www.w3.org/2001/XMLSchema#> .

<https://kb.virtualtreasury.ie/alignment> a align:Alignment ;
    align:onto1 <https://kb.virtualtreasury.ie/ontology> ;
    align:onto2 foaf: ;
    align:map [ a align:Cell ;
                align:entity1 <http://erlangen-crm.org/current/E21_Person> ;
                align:entity2 foaf:Person ;
                align:relation "=" ] .
<< <https://kb.virtualtreasury.ie/alignment>
   rdf:type align:Alignment >>
    metag:purpose       "Align person entities between the VRTI Knowledge Graph
                         and external datasets to enable integrated access." ;
    metag:mappingType   "Ontology Alignment" ;
    metag:mappingDomain "Historical" ;
    metag:startDate     "2024-03-01"^^xsd:date ;
    metag:endDate       "2024-03-30"^^xsd:date ;
    metag:mappingMethod "Automatic" ;
    metag:mappingURI    <http://virtualrecordtreasury.ie/person_alignment_v1.0>
;
    metag:testingType   "Validation" ;
    metag:testingResult "Successful" ;
    metag:publisherName "VRTI team" ;
    metag:versionNumber "1.0" ;
    metag:versionDateTime "2024-03-30"^^xsd:date .
```

Listing 4.5 Named Graph annotation for an ontology alignment

```
@prefix metag: <http://example.com/metag/> .
@prefix xsd:   <http://www.w3.org/2001/XMLSchema#> .

metag:subject
    metag:purpose       "Align person entities between the VRTI Knowledge Graph
                         and external datasets to enable integrated access." ;
    metag:mappingType   "Ontology Alignment" ;
    metag:mappingDomain "Historical" ;
    metag:startDate     "2024-03-01"^^xsd:date ;
    metag:endDate       "2024-03-30"^^xsd:date ;
    metag:mappingMethod "Automatic" ;
    metag:mappingURI    <http://virtualrecordtreasury.ie/person_alignment_v1.0>
  ;
    metag:testingType   "Validation" ;
    metag:testingResult "Successful" ;
    metag:publisherName "VRTI team" ;
    metag:versionNumber "1.0" ;
    metag:versionDateTime "2024-03-30"^^xsd:date .
```

Interlinking

For interlinking specifications, no native RDF resource exists in the SPARQL query file. MetaSEMAP mints a new resource of type *InterlinkingOperation* to serve as the annotation subject. Listing 4.6 shows the RDF-star annotation for an interlinking specification. Listing 4.7 shows the Named Graph equivalent for the interlinking specification.

Listing 4.6 RDF-star annotation for an interlinking specification

```
@prefix owl:   <http://www.w3.org/2002/07/owl#> .
@prefix metag: <http://example.com/metag/> .
@prefix xsd:   <http://www.w3.org/2001/XMLSchema#> .

<http://example.com/interlink/interlinking.rq>
    a <http://example.com/ontology#InterlinkingOperation> ;
    <http://example.com/queryContent> """
        INSERT { ?vrtEntity owl:sameAs ?wikidataEntity . }
        WHERE  { ?vrtEntity rdf:type cidoc:E21_Person ;
                            foaf:name ?name .
                 ?wikidataEntity wdt:P31 wd:Q5 ;
                                 rdfs:label ?name .
                 FILTER(LANG(?name) = "en") } """ .
<< <http://example.com/interlink/interlinking.rq>
   rdf:type <http://example.com/ontology#InterlinkingOperation> >>
    metag:purpose "Connect Irish historical figures in VRTI to corresponding
Wikidata entries using owl:sameAs relationships." ;
    metag:mappingType   "Interlinking" ;
    metag:mappingDomain "Historical" ;
    metag:startDate     "2024-04-01"^^xsd:date ;
    metag:endDate       "2024-04-20"^^xsd:date ;
    metag:mappingMethod "Manual" ;
    metag:mappingURI    <http://virtualrecordtreasury.ie/interlinking_v1.0> ;
    metag:testingType   "Validation" ;
    metag:testingResult "Successful" ;
    metag:publisherName "VRTI team" ;
    metag:versionNumber "1.0" ;
    metag:versionDateTime "2024-04-20"^^xsd:date .
```

Listing 4.7 Named Graph annotation for an interlinking specification

```
@prefix metag: <http://example.com/metag/> .
@prefix xsd:   <http://www.w3.org/2001/XMLSchema#> .

metag:subject
    metag:purpose       "Connect Irish historical figures in VRTI to
corresponding
                         Wikidata entries using owl:sameAs relationships." ;
    metag:mappingType   "Interlinking" ;
    metag:mappingDomain "Historical" ;
    metag:startDate     "2024-04-01"^^xsd:date ;
    metag:endDate       "2024-04-20"^^xsd:date ;
    metag:mappingMethod "Manual" ;
    metag:mappingURI    <http://virtualrecordtreasury.ie/interlinking_v1.0> ;
    metag:testingType   "Validation" ;
    metag:testingResult "Successful" ;
    metag:publisherName "VRTI team" ;
    metag:versionNumber "1.0" ;
    metag:versionDateTime "2024-04-20"^^xsd:date .
```

Comparing Listings 4.2, 4.4, 4.6 with Listings 4.3, 4.5, 4.7 illustrates the structural difference between the two representations. The RDF-star output embeds metadata

inline within the mapping file using the <<...>> syntax, keeping both the mapping logic and its contextual metadata in a single document. The Named Graph output produces a standalone metadata file, separating metadata from the mapping specification. The metadata fields are identical in both cases. As discussed in Study 2 (Section 5.4), participants in the usability evaluation showed a clear preference for the Named Graph representation, which is consistent with the readability benefit of keeping metadata as a separate, independently accessible artefact. The usability and effectiveness of these representations are discussed in full in Chapter 5, Study 2.

4.8 Evolution of the Model Based on User and Expert Feedback

The design of LMMD evolved iteratively through a series of publications, prototype implementations, and evaluation studies. This section describes how the model developed from an initial conceptual proposal to a formally implemented ontology, and summarises the feedback received through evaluation and the refinements applied to the final version.

4.8.1 Model Development and Initial Assumptions

The initial conceptual exploration of a unified metadata model for declarative mappings was presented at the Ontology Matching Workshop in 2022 [66]. At this stage, the model proposed five lifecycle phases including analysis, design, development, testing, and maintenance, with grouped metadata categories for each phase, as shown in Table 4.4.

Three core assumptions underpinned this initial design. First, that metadata could be captured at a conceptual level using grouped field categories without requiring individual property definitions. Second, that a single model structure would be applicable across all mapping types at project level. Third, that string-valued

properties would be sufficient for capturing contextual metadata such as purpose, domain, and stakeholder information.

Study 1, a community survey on metadata requirements [65], was conducted to evaluate whether the proposed lifecycle-based metadata fields were perceived by practitioners as relevant for mapping discovery and quality assessment tasks. The results confirmed the relevance of the proposed fields across both tasks, and qualitative responses suggested several additional metadata elements that informed subsequent formalisation decisions. These findings provided the foundational evidence needed to proceed with formalising the model.

Based on the confirmed field relevance, a tool called MetaMap was introduced at IEEE ICSC 2024 [70] to demonstrate the feasibility of lifecycle-based metadata annotation in practice. MetaMap focused on uplift mappings and served as an early proof of concept. Building on this, MetaSEMAP was developed as an improved tool supporting all three mapping types. MetaSEMAP was implemented using a simple prefix-based vocabulary rather than a full OWL ontology, reflecting the practical need to deploy and evaluate the tool before the ontology formalisation was complete. The model was subsequently formalised as the Mapping Metadata Vocabulary (MMV), an OWL 2 DL ontology, as described in Section 4.6 [83]. Study 1 also confirmed coverage of 34 out of 37 proposed metadata fields in MMV v0.1. Three fields which are testing identifier, testing name, and testing date/time were rated as lower priority and are noted as future work. Two stakeholder fields including URI and organisation were rated highly relevant and were subsequently addressed in MMV v0.2 through foaf:homepage and foaf:organization on the stakeholder agent.

4.8.2 Evaluation Feedback Refinements

Study 2, the usability evaluation of MetaSEMAP [71], showed that users were able to complete lifecycle-based metadata annotation tasks across different mapping types with generally positive usability results. The evaluation highlighted that free-text fields required additional guidance to promote consistency. A clear preference

was observed for the Named Graph representation over RDF-star under the tested conditions, informing the representation choices discussed in Section 4.7.3. These refinment were considered and applied in the next version of the tool.

Study 3, the expert evaluation of MMV, provided the most targeted feedback on the ontology design. Experts generally viewed MMV as conceptually clear, structurally coherent, and aligned with established ontology engineering practices, with strong support for the clarity of classes and properties and for vocabulary reuse. However, the evaluation also identified several areas for refinement. In particular, experts noted that string-valued properties for concept-oriented metadata such as mapping purpose, domain, and stakeholder information would benefit from being modelled as object properties linked to controlled vocabularies, improving semantic precision and interoperability. Experts also indicated that purpose-related metadata should evolve toward a more formally structured representation, and requested richer documentation examples and SHACL-based validation guidance to support correct instantiation of the model.

Study 4, the mapping reuse experiment, provided further insight into which metadata elements are most critical for reuse decisions and where additional guidance could strengthen the model in practice.

4.9 Chapter Summary

This chapter presented the design and implementation of the Lifecycle-Based Metadata Model for Declarative Mappings (LMMD). The chapter began by identifying three recurring challenges in current declarative mapping practice: type-specific documentation, artefact-centric metadata limitations, and limited lifecycle and validation support, which directly correspond to the gaps identified in Chapter 3. Based on these challenges, a domain analysis, and guiding considerations from FAIR principles, seven design requirements were derived

To address these requirements, LMMD was designed around six core features: a project-centric representation, lifecycle-based organisation, separation of activities and artefacts, a mapping-type agnostic core, standards reuse and alignment, and dual human-machine operationalisation. These features collectively ensure that LMMD provides structured, reuse-oriented metadata applicable across uplift mappings, interlinking specifications, and ontology alignment. Table 4.3 in this chapter summarises how each design feature addresses the seven requirements defined in Section 4.4.

The model was formalised as the Mapping Metadata Vocabulary (MMV), an OWL 2 DL ontology that introduces 9 new classes, 16 new object properties, and 27 new data properties to capture lifecycle-specific concepts not adequately covered by existing standards such as PROV-O, Dublin Core Terms, FOAF, DQV, and MQV. The ontology was developed following the LOT and NeOn methodologies.

LMMD is operationalised through MetaSEMAP, a guided web-based annotation tool that enables lifecycle-based metadata annotation across different mapping types without requiring users to interact directly with ontology languages. The model evolved from an initial conceptual proposal in 2022 through iterative evaluation, as described in Section 4.8.

The evaluation of LMMD across four complementary studies is presented in Chapter 5, with an additional use case exploring AI-assisted mapping generation presented in Chapter 6.

Chapter 5: Evaluation

This chapter presents the findings of the four studies undertaken to evaluate the Lifecycle-Based Metadata Model for Declarative Mappings (LMMD). The four studies examine LMMD from complementary perspectives across different instruments and metrics. These include community requirements, usability in practice, ontology quality, and support for mapping reuse.

A summary of each study is provided in Section 5.1. Subsequent sections present each study and its findings in detail. A chapter summary of the evaluations is provided in Section 5.8.

5.1 Overview of Evaluation Strategy

This section provides an overview of the four studies conducted as part of the evaluation of the Lifecycle-Based Metadata Model for Declarative Mappings (LMMD). Each study addresses a specific aspect of lifecycle-based metadata and, together they form a multi-method evaluation strategy that examines LMMD from both conceptual and practical perspectives. An overview of the four studies and their focus areas is shown in **Figure 5.1**.

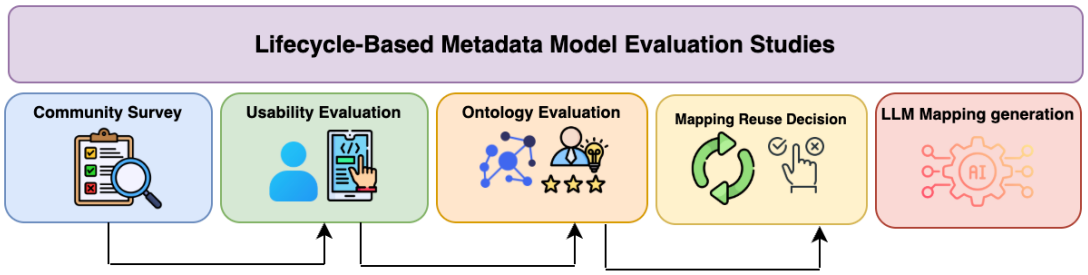


Figure 5.1 Overview of the four evaluation studies conducted to assess LMMD

- **Study 1** investigated community perspectives on metadata requirements for declarative mappings. A survey-based approach was used to examine which

metadata elements practitioners consider important for mapping tasks including discovery and quality assessment.

- **Study 2** evaluated the usability of MetaSEMAP, a tool implementing LMMD. A controlled user study was used to assess the ability of the users to annotate mappings using lifecycle metadata, as well as their experiences with alternative metadata representations.
- **Study 3** evaluated the MMV ontology. This study used automated validation tools and expert evaluation to assess whether the ontology meets its intended requirements and complies with established ontology design principles. Study 3 also examined whether MMV correctly answers the competency questions defined in Section 4.5.3, providing formal validation of the ontology's coverage.
- **Study 4** investigated the role of lifecycle metadata in supporting mapping reuse decisions. Participants compared two metadata representations and provided judgements on reuse decisions, representation preference, understandability, and decision confidence.
- A use case study presented in Chapter 6 explored the use of lifecycle metadata in guiding LLM-based mapping generation, demonstrating a potential application of LMMD beyond its original evaluation scope

The evaluation studies presented in this chapter build on the iterative development of the lifecycle-based metadata model throughout this research. The foundational work on a unified metadata model for semantic and data mappings was presented at the Ontology Matching Workshop (OM 2022)[66]. The community survey on metadata requirements, detailed in Study 1, was published at IEEE ICSC 2023 [65]. An initial prototype tool, MetaMap, was introduced at IEEE ICSC 2024 [70], which motivated further refinement of the tool. The usability evaluation of the improved tool, MetaSEMAP (Study 2), was presented at the UKG 2025 workshop co-located with the SEMANTiCS 2025 conference [71]. The ontology formalisation of the model (MMV), evaluated in Study 3, was presented at the SDI Workshop at IEEE ICSC 2026 [83]. The use case study on metadata-guided LLM

mapping generation, which supports the minor contribution of this thesis, was presented at the KG-NeSy 2025 workshop [72], also at SEMANTiCS 2025. The full study is presented in Chapter 6. Feedback received at these venues informed subsequent refinements of the model, ontology, and tool.

Across the four evaluation studies, participants are described using two terms. The term *practitioner* refers to anyone with experience working with declarative mappings regardless of professional role, including MSc students in knowledge engineering, researchers, and data engineers. The term *expert* is used in two distinct senses: in Study 1, it refers to participants who self-reported expert-level experience with declarative mappings on the recruitment questionnaire; in Study 3, it refers to ontology engineers and Semantic Web researchers recruited on the basis of prior domain expertise. Table 5.1 provides an overview of participant profiles across the four studies.

Table 5.1 Participant profiles across evaluation studies

Study	n	Term Used	Role	Experience / Background
Study 1 Community Survey	67	Practitioners	Researchers and practitioners (self-selected)	Expert (9), Intermediate (9), Beginner (49)
Study 2 Usability Evaluation	46	Practitioners	MSc students (Knowledge and Data Engineering, TCD)	Introductory knowledge of declarative mappings, varied technical experience
Study 3 Expert Evaluation	6	Experts	Ontology engineers and Semantic Web researchers	Prior experience in ontology development and Semantic Web research (inclusion criteria applied)
Study 4 Reuse Experiment	10	Practitioners and Experts	Researchers and practitioners in declarative mappings	Recruited via semantic-web@w3.org and direct outreach

The four studies discussed in this chapter are numbered for reference purposes only. The numbering does not imply a sequential dependency between studies, each addresses a distinct evaluation question and was conducted independently. However, the studies share a common empirical foundation. The Community Survey (Study 1) established which metadata fields practitioners consider relevant

for mapping discovery and quality assessment. The subsequent studies build on this foundation: Study 2 assessed whether those fields could be effectively annotated in practice; Study 3 examined whether MMV formally represents them correctly; and Study 4 tested whether they support informed reuse decisions. Chapter 6 extends this foundation by investigating whether the same fields can guide automated mapping generation as use case of the model.

5.2 Evaluation instruments

The instruments used across the four studies are summarised in this section and described in detail within each study. These instruments were selected to support a multi-method evaluation of LMMD, capturing community perspectives, user experience, conceptual quality, and exploratory impacts on reuse and automated mapping generation.

Figure 5.2 summarises the evaluation instruments and associated measures used across the four studies. The figure is organised by study, highlighting the instruments applied in each case and the primary constructs measured, including relevance, usability, ontology quality, reuse decision support, and comparative quality of LLM-generated mappings.

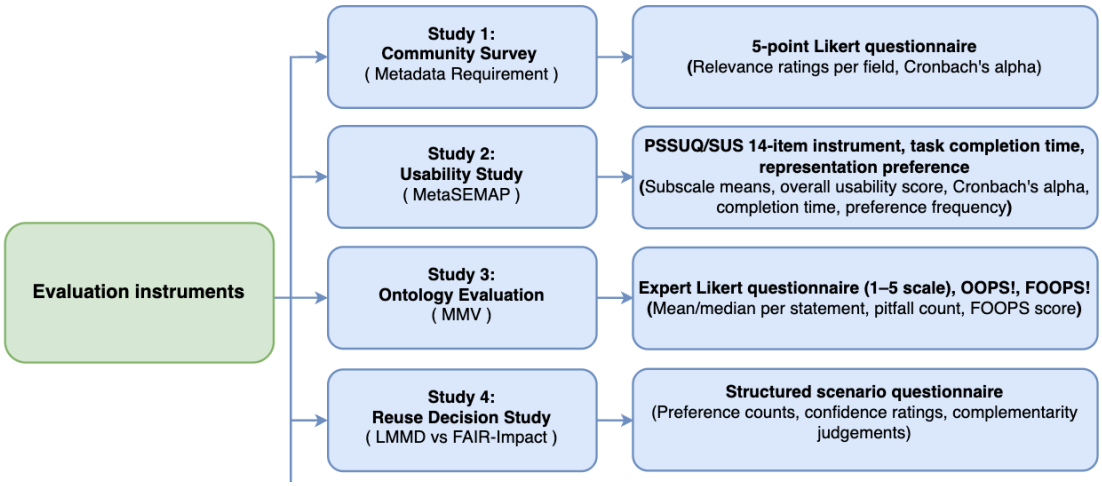


Figure 5.2 Evaluation instruments and measures across four studies

5.3 Study 1: Community Survey on Metadata Requirements

This section presents the first study conducted to evaluate the Lifecycle-Based Metadata Model for Declarative Mappings (LMMD). Study 1 is a community survey designed to assess whether metadata fields proposed in LMMD are perceived as relevant by the practitioners working with declarative mappings. The study was conducted in two rounds: an initial round distributed between March and July 2022 with 18 participants and a follow-up round that remained open until November 2022 with 49 participants completing the survey, resulting in a combined dataset of 67 participants. The initial round results were published at IEEE ICSC 2023 [65]. The section is structured into the following subsections: Motivation; Hypotheses; Methodology; Participants and recruitment; Results; and Discussion.

5.3.1 Motivation

Declarative mapping artefacts are commonly shared as standalone technical files with little or no accompanying documentation. While such artefacts may be executable, the absence of structured contextual metadata makes it difficult for third parties to discover suitable mappings, interpret their intent, or assess their quality and suitability for reuse. This problem was confirmed by the challenges reported in Chapter 4, particularly those related to artefact-centric documentation and limited lifecycle coverage.

As described in Section 4.5.3, the design of LMMD was guided by those competency questions created by the author of this thesis as a validation instrument during development. While this approach ensured that the model addressed the requirements derived from the state-of-the-art analysis, internally authored competency questions reflect the researcher's own synthesis of the literature and assumptions about user needs. External validation from practitioners who work with declarative mappings in practice was therefore considered useful to confirm whether the proposed metadata fields would address genuine information needs of

the community rather than assumed ones. The 37 metadata fields evaluated in Study 1 correspond directly to the properties defined in LMMD and capture the same constructs targeted by the competency questions described in Section 4.5.3. Evaluating the perceived relevance of these fields therefore also provides indirect empirical support for the competency questions that motivated them.

Therefore, Study 1 was designed to provide external validation. Practitioners were asked to rate the relevance of each proposed metadata field for two key mapping-related tasks: discovering mappings suitable for reuse, and assessing the quality and trustworthiness of existing mappings. These two tasks were selected because they represent the primary use cases motivating LMMD. Mapping discovery requires metadata describing the context, purpose, and provenance of a mapping so that a practitioner can judge suitability without inspecting the mapping directly. Quality assessment requires metadata documenting how a mapping was developed, tested, and maintained so that reliability can be judged. The results of Study 1 directly informed subsequent formalisation decisions in MMV, as discussed in Chapter 4.

5.3.2 Hypothesis

The following hypotheses were defined for Study 1:

- **Hypothesis 1 (Discovery relevance):** The lifecycle-based metadata fields proposed in LMMD are perceived as relevant by practitioners for the task of discovering declarative mappings suitable for reuse.
- **Hypothesis 2 (Quality relevance):** The lifecycle-based metadata fields proposed in LMMD are perceived as relevant by practitioners for the task of assessing the quality and trustworthiness of declarative mappings.
- **Hypothesis 3 (Experience effect):** Participants with greater experience in declarative mappings assign higher relevance to the proposed metadata fields than participants with less experience.

5.3.3 Methodology

A questionnaire-based survey was conducted to test the hypotheses. Questionnaire-based surveys are well established in requirements engineering and early-stage model validation research, particularly for formative evaluation of proposed models prior to full implementation [73]. This approach was appropriate given that Study 1 is positioned as a formative evaluation intended to validate requirements before full ontology formalisation and tool deployment. The survey was conducted individually and anonymously, without a time limit. Participants were asked to rate the relevance of each proposed metadata field for two specific mapping-related tasks. Both quantitative and qualitative data were collected. Quantitative data consisted of five-point Likert-scale relevance ratings for each of the 37 proposed metadata fields where 1 denotes Highly Relevant and 5 denotes Irrelevant. Qualitative data consisted of open-ended responses allowing participants to suggest additional metadata fields or comment on the proposed model.

Ethical approval was received from the Research Ethics Committee at Trinity College Dublin. The study materials are provided in **Appendix D** (ethics approval, invitation email, consent form, survey instrument, and Round 2 tables). Raw data and study materials are also available at https://github.com/SarahAlzahranitcd/Thesis-Materials/tree/main/Study-1-Community-Survey.

5.3.4 Evaluation instruments

The following measures were used to analyse the survey responses.

Internal consistency. Cronbach's alpha was computed separately for Task 1 and Task 2 to assess whether participants responded to the 37 metadata fields in a consistent manner [74].

Rating distributions. For each metadata field, the number of participants assigning each rating (1-5) is reported per experience group. This count-based approach preserves the full distribution of responses and is consistent with the approach used in the initial publication of Study 1 [65]. Results are presented in Tables 5.2- 5.5.

Acceptance criterion. A metadata field is considered to have received positive relevance support if its median rating is ≤ 2 among expert and intermediate participants combined, where 1 denotes Highly Relevant and 5 denotes Irrelevant. H1 and H2 are considered supported if the majority of the 37 proposed fields meet this criterion.

Phase-level means and standard deviations. Mean ratings and standard deviations were computed at the lifecycle phase level, aggregating all fields belonging to the same phase. This summarises patterns across the 37 fields in a way that reflects the lifecycle-oriented structure of LMMD. Lower mean values indicate higher perceived relevance. Results are presented in Table 5.6.

5.3.5 Experiment Structure

The survey was structured as follows:

1. **Informed consent:** At the beginning of the survey, participants were provided with an information sheet for additional details regarding the study. Participants were required to indicate their consent before proceeding. The participant information sheet and consent form are provided in Appendix D.
2. **Questionnaire-based survey (Likert scale)**: Participants were asked to rate the relevance of individual metadata fields on a five-point Likert scale ranging from *Highly Relevant(1) to Irrelevant(5)*. The questionnaire covered metadata fields that are associated with all lifecycle phases represented in LMMD. Each metadata field was evaluated separately for two mapping-related tasks: mapping discovery for reuse, and mapping quality assessment. For example,

participants were asked to rate how relevant a field such as mapping domain to the task of discovering a mapping for reuse.

3. **Open-ended questions**: At the end of the survey participants were asked to suggest additional metadata elements they considered missing from the model. They also were invited to comment on the scope and clarity, and provide feedback on the completeness of the lifecycle-based metadata model.

5.3.6 Participants and Recruitment

Participation invitations were distributed via community mailing lists (including *semantic-web@w3.org*), and through direct outreach to researchers and practitioners working on declarative mappings and related topics. This sampling strategy reflected the target population for LMMD: practitioners familiar with mapping creation, use, and evaluation in Semantic Web and Linked Data workflows.

Participants self-reported their experience with declarative semantic and data mappings by selecting one of four levels: no experience, beginner, intermediate, or expert. Participants who selected no experience were excluded from all analyses, as they cannot reliably assess the relevance of metadata fields for mapping-related tasks.

The study was conducted in two rounds. The initial round was distributed between March and July 2022 yielded 18 participants consisting of 9 experts, 6 intermediates, and 3 beginners. These results were published at IEEE ICSC 2023 [65]. Following publication, the survey remained open and a follow-up round collected an additional 49 responses between August and November 2022, comprising 3 intermediates and 46 beginners. The combined dataset totals 67 participants.

For the initial round, three participants did not complete the quality assessment task (Task 2): two experts and one beginner. These participants are included in the

Task 1 analysis but excluded from the Task 2 analysis. Task 1 was therefore completed by 18 participants (9 experts, 6 intermediates, 3 beginners) and Task 2 by 16 participants (8 experts, 6 intermediates, 2 beginners). Table 5.2 summarises the participant breakdown across both rounds.

Table 5.2 Participant breakdown by round and experience level

Round	Expert	Intermediate	Beginner	Total
Initial round (Mar-Jul 2022)	9	6	3	18
Follow-up round (Aug-Nov 2022)	0	3	46	49
Combined	9	9	49	67

5.3.7 Results

This section presents the results of Study 1 across both rounds. Section 5.3.7.1 presents the initial round results (n=18), which formed the basis of the published analysis [65]. Section 5.3.7.2 presents the follow-up round results (n=49) and compares findings across rounds.

The survey instrument showed excellent reliability for both tasks (Cronbach's α = 0.905 for Task 1; α = 0.910 for Task 2). This indicates that participants who rated one field as relevant tended to rate related fields as relevant, suggesting the 37 metadata fields form a reliable instrument.

5.3.7.1 Initial Round Results

The following subsections present the rating distributions for each task, followed by phase-level summary statistics.

Task 1: Mapping Discovery for Reuse

Tables 5.3 and 5.4 present the rating distributions for Task 1 by expert and intermediate participants respectively. For each field, the number of participants assigning each rating (1-5) is shown. In Table 5.3, for example, four experts rated

the Stakeholder Name field as relevant (combining ratings 1 and 2), two rated it as neutral, one as less relevant, and two as irrelevant. Five fields were proposed to describe stakeholders: name, URI, role, background, and organisation.

Table 5.3 Ratings results of the metadata fields relevance to mapping discovery task by Experts

Mapping discovery task - Experts						
	Proposed fields	**1 Highly relevant**	**2**	**3**	**4**	**5 irrelevant**
Metadata fields describing stakeholders	Name	2	2	2	1	2
	URI	3	1	1	3	1
	Role	2	3	3	1	0
	Background	1	1	3	4	0
	organization	2	1	5	1	0
Metadata fields describing the mapping project's purpose	Requirement	5	3	0	0	0
	Type of mapping	6	3	0	0	0
	Domain	4	3	2	0	0
	assumptions	2	4	3	0	0
	Technical Requirement	2	3	4	0	0
	Risks/issues	1	4	2	2	0
Metadata fields describing the input to be mapped	Name	1	4	2	1	1
	URI	5	1	1	1	1
	Source	5	1	2	1	0
	Type	7	0	2	0	0
	Creator	1	3	1	3	1
	Format	4	1	3	0	1
Metadata fields describing the mapping design process	Design decisions	4	2	2	0	1
	Justifications	3	3	2	1	0
	Quality Metrics	4	3	2	0	0
Metadata fields describing the mapping development process	Mapping Name	2	3	2	0	2
	Mapping URI	6	0	2	0	1
	Start date and time	1	3	2	1	2
	End date and time	1	4	2	1	1
	Used tools	3	1	4	0	1
	Mapping method	5	3	1	0	0
	Mapping algorithm	3	3	2	1	0
	Mapping format	6	2	1	0	0
Metadata fields describing the testing process of a mapping	Testing Identifier	3	0	3	1	2
	Testing Name	1	2	4	2	0
	Testing type	3	3	3	0	0
	Testing date and time	1	2	3	2	1
	Testing result	4	3	2	0	0
Metadata fields describing the management and maintenance of mappings	Publisher name	4	3	1	0	1
	Publisher source	5	0	3	0	1
	Version number	5	2	0	1	1
	Version date/time	5	3	0	0	1

Table 5.4 Ratings results of the metadata fields relevance to mapping discovery task by Intermediates

Mapping discovery task – intermediates						
	Proposed fields	**1 Highly relevant**	**2**	**3**	**4**	**5 irrelevant**
Metadata fields describing stakeholders	Name	2	3	0	1	0
	URI	2	2	1	1	0
	Role	1	3	0	2	0
	Background	1	3	0	2	0
	organization	1	2	2	1	0
Metadata fields describing the mapping project's purpose	Requirement	4	1	0	1	0
	Type of mapping	4	1	1	0	0
	Domain	5	1	0	0	0
	assumptions	2	3	0	1	0
	Technical Requirement	2	2	0	2	0
	Risks/issues	2	2	0	2	0
Metadata fields describing the input to be mapped	Name	3	3	0	0	0
	URI	4	2	0	0	0
	Source	4	1	1	0	0
	Type	3	1	2	0	0
	Creator	1	2	2	1	0
	Format	1	2	3	0	0
Metadata fields describing the mapping design process	Design decisions	0	3	3	0	0
	Justifications	0	3	3	0	0
	Quality Metrics	0	4	2	0	0
Metadata fields describing the mapping development process	Mapping Name	2	3	0	0	1
	Mapping URI	2	3	0	0	1
	Start date and time	1	0	2	2	1
	End date and time	2	0	1	2	1
	Used tools	1	3	0	2	0
	Mapping method	1	4	1	0	0
	Mapping algorithm	1	2	3	0	0
	Mapping format	1	1	4	0	0
Metadata fields describing the testing process of a mapping	Testing Identifier	2	2	0	2	0
	Testing Name	1	1	2	2	0
	Testing type	3	2	0	1	0
	Testing date and time	1	1	1	2	1
	Testing result	3	2	1	0	0
Metadata fields describing the management and maintenance of mappings	Publisher name	2	2	2	0	0
	Publisher source	3	1	2	0	0
	Version number	1	3	2	0	0
	Version date/time	3	0	3	0	0

Six experts provided open-ended responses suggesting additional fields for the discovery task. Two experts pointed to the Simple Standard for Sharing Ontology Mappings (SSSOM) [14]. as a complementary standard, specifically recommending consideration of mapping cardinality, mapping justification, and confidence fields. Additional suggestions included output metadata describing the target of the mapping, usage metrics such as number of projects and production use, mapping constructs, and change or issue management information. These suggestions informed subsequent refinements to LMMD and MMV, as discussed in Chapter 4.

Task 2: Mapping Quality Assessment

Tables 5.5 and 5.6 present the rating distributions for Task 2 by expert and intermediate participants respectively. Task 2 was completed by 16 participants (8 experts, 6 intermediates, 2 beginners), as three participants did not complete this section.

One expert proposed additional fields for quality assessment from SSSOM [14], including mapping cardinality, mapping justification, and mapping confidence. One expert noted that the Mapping Algorithm field required clarification, as it was unclear whether it referred to the algorithm used to create the mapping or to execute it. One expert indicated no distinction between the two tasks, noting that the same fields are relevant to both discovery and quality assessment.

Table 5.5 Rating results of the metadata fields relevance to mapping quality assessment task by Experts

Mapping quality assessment task - Experts						
	Proposed fields	**1 Highly relevant**	**2**	**3**	**4**	**5 irrelevant**
Metadata fields describing stakeholders	Name	1	3	1	2	1
	URI	2	1	2	2	1
	Role	2	2	3	1	0
	Background	1	1	4	2	0
	organization	2	3	2	1	0
Metadata fields describing the mapping project's purpose	Requirement	7	1	0	0	0
	Type of mapping	6	1	0	1	0
	Domain	4	2	2	0	0
	assumptions	3	5	0	0	0
	Technical Requirement	3	3	1	1	0
	Risks/issues	1	4	1	2	0
Metadata fields describing the input to be mapped	Name	0	3	2	1	2
	URI	2	2	1	2	1
	Source	4	1	1	2	0
	Type	5	1	0	2	0
	Creator	1	2	0	4	1
	Format	3	1	1	2	1
Metadata fields describing the mapping design process	Design decisions	2	5	0	0	1
	Justifications	3	3	2	0	0
	Quality Metrics	4	4	0	0	0
Metadata fields describing the mapping development process	Mapping Name	0	3	1	2	2
	Mapping URI	4	1	1	1	1
	Start date and time	1	2	1	2	2
	End date and time	1	3	1	2	1
	Used tools	4	0	2	1	1
	Mapping method	4	4	0	0	0
	Mapping algorithm	3	3	2	0	0
	Mapping format	2	2	1	1	2
Metadata fields describing the testing process of a mapping	Testing Identifier	2	1	3	0	2
	Testing Name	0	3	3	0	2
	Testing type	5	1	2	0	0
	Testing date and time	1	4	1	1	2
	Testing result	5	2	1	0	0
Metadata fields describing the management and maintenance of mappings	Publisher name	3	1	1	1	2
	Publisher source	2	0	4	1	1
	Version number	3	2	0	2	1
	Version date/time	3	3	0	0	2

Table 5.6 Ratings results of the metadata fields relevance to mapping quality assessment task by intermediates

Mapping quality assessment task – intermediates						
	Proposed fields	**1 Highly relevant**	**2**	**3**	**4**	**5 irrelevant**
Metadata fields describing stakeholders	Name	1	3	1	1	0
	URI	0	3	2	1	0
	Role	1	1	2	2	0
	Background	1	4	1	0	0
	organization	1	2	0	3	0
Metadata fields describing the mapping project's purpose	Requirement	2	3	0	1	0
	Type of mapping	2	3	0	1	0
	Domain	3	1	0	1	1
	assumptions	3	3	0	0	0
	Technical Requirement	2	3	0	1	0
	Risks/issues	2	2	2	0	0
Metadata fields describing the input to be mapped	Name	2	4	0	0	0
	URI	3	3	0	0	0
	Source	5	0	0	1	0
	Type	3	2	0	1	0
	Creator	2	1	2	1	0
	Format	1	4	0	1	0
Metadata fields describing the mapping design process	Design decisions	4	1	1	0	0
	Justifications	4	1	1	0	0
	Quality Metrics	4	1	1	0	0
Metadata fields describing the mapping development process	Mapping Name	2	2	2	0	0
	Mapping URI	2	2	2	0	0
	Start date and time	1	2	0	2	1
	End date and time	1	2	0	2	1
	Used tools	2	1	2	1	0
	Mapping method	5	0	1	0	0
	Mapping algorithm	3	2	1	0	0
	Mapping format	1	2	2	1	0
Metadata fields describing the testing process of a mapping	Testing Identifier	4	0	2	0	0
	Testing Name	3	0	3	0	0
	Testing type	5	0	1	0	0
	Testing date and time	4	0	1	1	0
	Testing result	5	1	0	0	0
Metadata fields describing the management and maintenance of mappings	Publisher name	2	3	1	0	0
	Publisher source	3	1	0	2	0
	Version number	1	4	0	1	0
	Version date/time	3	1	0	2	0

Table 5.7 presents the phase-level mean ratings and standard deviations for the initial round. For each phase, the mean was computed by averaging all field ratings from participants within that experience group, and the standard deviation indicates the degree of agreement among participants within the group. Lower mean values indicate higher perceived relevance. The Analysis-Purpose metadata fields received the lowest mean ratings from both expert and intermediate participants across both tasks, indicating the strongest perceived relevance. The Analysis-Stakeholder metadata fields showed the highest variability, particularly among beginner participants in Task 2 (mean = 4.50, SD = 0.50), suggesting that less experienced practitioners do not recognise the value of stakeholder metadata when assessing mapping quality. The Maintenance and Design phases were consistently rated as relevant by expert and intermediate participants across both tasks.

Table 5.7 Phase-level mean ratings and standard deviations for the initial round by experience group

Lifecycle Phase	Expert (n=9)	Intermediate (n=6)	Beginner (n=3)
	Mean - Standard Deviation		
Task 1: Mapping Discovery for Reuse			
Analysis (Stakeholders)	2.73 - 1.22	2.33 - 1.07	3.80 - 0.98
Analysis (Purpose)	1.91 - 0.85	1.86 - 1.11	2.39 - 1.21
Analysis (Inputs)	2.22 - 1.30	1.83 - 0.90	2.39 - 1.06
Design	2.00 - 1.05	2.44 - 0.50	2.56 - 1.50
Development	2.22 - 1.26	2.50 - 1.21	3.04 - 1.17
Testing	2.49 - 1.19	2.37 - 1.25	3.53 - 1.15
Maintenance	1.97 - 1.32	2.00 - 0.87	2.67 - 1.31
Task 2: Mapping Quality Assessment			
Analysis (Stakeholders)	2.57 - 1.18	2.53 - 0.99	4.50 - 0.50
Analysis (Purpose)	1.69 - 0.89	1.97 - 1.07	1.92 - 1.11
Analysis (Inputs)	2.64 - 1.43	1.83 - 0.96	2.00 - 1.29
Design	1.86 - 0.94	1.50 - 0.76	1.67 - 0.94
Development	2.54 - 1.45	2.23 - 1.18	2.33 - 1.40
Testing	2.46 - 1.34	1.60 - 0.95	1.90 - 0.94
Maintenance	2.64 - 1.56	2.08 - 1.11	2.12 - 1.27

5.3.7.2 Follow-up Round Results

The follow-up round collected 49 additional responses between August and November 2022, comprising 3 intermediates and 46 beginners. The full rating distributions for the follow-up round are provided in **Appendix D** (Tables D.1–D.4).

Table 5.8 presents phase-level mean ratings and standard deviations for both rounds side by side, allowing direct comparison across participant groups and rounds. The follow-up round results are consistent with the initial round findings, providing additional confidence that the perceived relevance of the proposed metadata fields generalises beyond the initial expert-oriented sample.

Table 5.8 Phase-level mean ratings and standard deviations for the follow-up round by experience group

Lifecycle Phase	Intermediate (n=3)	Beginner (n=46)
	Mean - Standard Deviation	
Task 1: Mapping Discovery for Reuse		
Analysis (Stakeholders)	1.93 - 0.77	2.46 - 1.18
Analysis (Purpose)	1.72 - 0.80	1.92 - 0.93
Analysis (Inputs)	1.83 - 0.50	2.15 - 1.04
Design	1.33 - 0.47	1.85 - 0.88
Development	1.29 - 0.54	2.22 - 1.05
Testing	1.53 - 0.62	2.21 - 1.11
Maintenance	1.25 - 0.43	2.18 - 1.10
Task 2: Mapping Quality Assessment		
Analysis (Stakeholders)	1.60 - 0.61	2.44 - 1.19
Analysis (Purpose)	1.39 - 0.59	1.95 - 1.02
Analysis (Inputs)	1.59 - 0.60	2.09 - 1.07
Design	1.44 - 0.50	1.87 - 0.95
Development	1.46 - 0.50	2.30 - 1.20
Testing	1.47 - 0.72	2.10 - 1.14
Maintenance	1.33 - 0.47	2.26 - 1.19

5.3.8 Discussion

This section discusses the results of Study 1 in relation to the three hypotheses defined in Section 5.3.2.

Hypothesis 1 and 2: Relevance of proposed metadata fields.

Both hypotheses are supported by the results. For Task 1 (mapping discovery), 29 out of 37 proposed metadata fields (78%) received positive relevance support, meeting the acceptance criterion of median ≤ 2 among expert and intermediate participants combined. For Task 2 (mapping quality assessment), 30 out of 37 fields (81%) met the same criterion. A one-sided binomial sign test confirmed that the proportion of fields meeting the acceptance criterion was significantly above chance for both H1 (78%, $p < 0.001$) and H2 (81%, $p < 0.001$), supporting the claim that the majority of proposed fields are perceived as relevant. In both cases the majority of the proposed fields were perceived as relevant, confirming that LMMD addresses genuine community information needs for both tasks.

The eight fields that did not meet the criterion in Task 1 share a common characteristic. They are administrative or procedural fields rather than conceptual ones. These include *Stakeholder Background, Stakeholder Organisation, Input Creator, Development Start Date, Development End Date, Testing Identifier, Testing Name*, and *Testing Date/Time*.

A similar pattern was observed in Task 2. This finding suggests that practitioners place higher value on metadata describing what a mapping does and why it was created, rather than metadata capturing who created it and when.

As a result of this study, fields without direct provenance value such as *Stakeholder Organisation, Testing Identifier*, and *Testing Name* are not represented as properties in MMV. Provenance-relevant fields such as *Development Start Date, Development End Date*, and *Testing Date/Time*, are represented via inherited PROV-O properties. This is consistent with MMV's design principle of reusing established vocabularies. Two lower-rated fields which

are *Stakeholder Background* and *Input Creator* were retained in MMV given their value for provenance-aware reuse.

The strongest relevance was consistently observed for the Analysis (Purpose) phase. *Mapping Requirements, Mapping Type*, and *Mapping Domain* metadata fields received median ratings of 1 across both tasks from expert and intermediate participants, indicating agreement on their relevance. This provides strong evidence that purpose-related metadata fields are central to both mapping tasks, supporting the primary design decision made in LMMD.

Hypothesis 3: Experience effect.

Hypothesis 3 is partially supported. Expert and intermediate participants rated the proposed fields as relevant across both tasks, with similar mean ratings (Task1: Expert= 2.23, Intermediate= 2.19). Beginner participants rated fields as less relevant overall (Task1: Beginner= 2.92), suggesting that experience with declarative mappings influences the perceived value of lifecycle metadata. Only experts suggested additional metadata fields in the open-ended responses, further supporting this pattern.

Analysis of Open-ended responses.

Six themes emerged from analysing the open-ended responses. Two experts recommended reviewing the SSSOM standard [14], specifically suggesting mapping cardinality, mapping justification, and confidence as additional fields. this suggestions of reusing SSSOM was out of the scope as this model is designed speicfically for ontology alignment and its focus on the correspondce level. One expert suggested output metadata describing the target of the mapping. One expert recommended usage metrics including number of projects and production use. One expert proposed mapping constructs as an additional field. One expert noted that

the *Mapping Algorithm* field was ambiguous, as it was unclear whether it referred to the algorithm used to create the mapping or to execute it. These suggestions were reviewed and informed subsequent refinements.

5.3.9 Experiment Summary

Study 1 provided foundational empirical evidence that lifecycle-based metadata is perceived by practitioners as relevant for both mapping discovery and mapping quality assessment. Hypothesis 1 and Hypothesis 2 were both supported, with 29 and 30 out of 37 fields respectively meeting the acceptance criterion. The results support the central claim of LMMD that reuse-oriented mapping documentation requires explicit representation of purpose, design rationale, validation activities, and maintenance context.

By evaluating the relevance of the proposed metadata fields before full ontology formalisation, Study 1 strengthened the methodological grounding of LMMD and informed subsequent design, ontology implementation, and tool-based evaluation. Limitations of the study include the modest sample size and reliance on self-reported expertise and relevance judgements. However, these limitations are mitigated by the expert-oriented participant pool where experts were contacted directly through their email in the initial round. The follow-up round results were consistent with the initial round, with the high rating on the group of metadata representing the purpose of the mapping, confirming that the relevance of purpose-related metadata generalises beyond the initial expert sample.

While Study 1 provided evidence that lifecycle-based metadata is perceived as relevant at a conceptual level, it did not examine whether such metadata can be effectively authored and managed in practice. This motivated Study 2, which evaluated the usability of MetaSEMAP as an implementation of LMMD.

5.4 Study 2: Usability Evaluation of MetaSEMAP

This section presents a usability evaluation of MetaSEMAP, a tool designed to facilitate the annotation of declarative mappings, including uplift mappings, ontology alignment, and interlinking. The study evaluated how users interact with the tool, how effectively they completed annotation tasks across different mapping types, and whether the choice of metadata representation influenced their interpretation and preference. The usability evaluation reported in this section was presented at the UKG 2025 workshop co-located with SEMANTiCS 2025 [71].

This section is structured as follows: Motivation; Hypothesis; Methodology; Evaluation Instruments; Experiment Structure; Participants and Recruitment; Results; Discussion; and Experiment Summary.

5.4.1 Motivation

Comprehensive and well-structured metadata annotation is essential for supporting the reuse and maintenance of declarative mappings throughout their lifecycle. The practical value of a lifecycle-based metadata model depends not only on its conceptual design but on whether users can engage with it effectively. A well-designed metadata structure may fail to achieve adoption if users find the annotation process complex, confusing, or cognitively demanding.

MetaSEMAP (described in Section 4.7) was developed as a web-based tool to facilitate structured annotation of uplift mappings, ontology alignment, and interlinking using the lifecycle-based metadata model defined in LMMD. The tool generates metadata in two representations: RDF-star, which embeds metadata inline within the mapping file, and Named Graph, which produces metadata as a separate file from the mapping specification. Evaluating how users interact with MetaSEMAP, how effectively they complete annotation tasks across different mapping types, and whether the choice of representation influences their interpretation was therefore a necessary step in assessing the practical feasibility of the model.

The primary objective of this study was to conduct an exploratory usability evaluation of MetaSEMAP with participants who have foundational but non-expert knowledge of mapping processes. The study examines whether lifecycle-based metadata annotation is feasible across all three mapping types, and whether separating metadata from the mapping file affects how users read and interpret the generated output.

5.4.2 Hypothesis

The following hypotheses were defined for this experiment:

- **Hypothesis 1 (Usability):** Participants will report positive usability ratings for MetaSEMAP when annotating declarative mappings using lifecycle-based metadata.
- **Hypothesis 2 (Task effectiveness):** Participants will be able to complete metadata annotation tasks across all three declarative mapping types supported by MetaSEMAP.
- **Hypothesis 3 (Representation preference):** Participants will express a clear preference for one metadata representation (Named Graph or RDF-star) when reviewing generated annotations.

  These hypotheses reflect the central implementation claim of LMMD that lifecycle-based metadata can be operationalised in a usable tool environment and that alternative metadata representations may influence user interpretation.

5.4.3 Methodology

A structured user evaluation of MetaSEMAP was conducted to test the hypotheses. Structured usability studies are commonly used in human-computer interaction research to assess user experience, task performance, and perceived effectiveness

of a tool [75]. This approach was appropriate for evaluating MetaSEMAP as a practical implementation of LMMD.

The experiment was carried out individually and anonymously through the public MetaSEMAP website (https://metasemap-1-sarah-alzahrani.replit.app/). At the time of this evaluation, MetaSEMAP generated metadata using the metag: prefix-based vocabulary, which was the implementation state prior to full MMV alignment. The metag: terms correspond directly to MMV properties, for example metag:purpose maps to mmv:hasPurpose. This is described in Section 4.7.3.

In this version of the tool, optional fields left blank by the user were serialised as empty string literals in the generated output. MetaSEMAP was developed iteratively, and this aspect of the output representation was refined in the next iteration of the tool, which omits unpopulated optional fields entirely so that no triple is created for blank values.

Each participant was randomly assigned one of three scenarios representing a different declarative mapping type: uplift mapping, ontology alignment, or interlinking. All scenarios were grounded in the Virtual Record Treasury of Ireland (VRTI)[7] knowledge graph. which provides structured, domain-rich data suitable for metadata annotation tasks.

Ethical approval was received from the Research Ethics Committee at Trinity College Dublin. The study materials are provided in **Appendix E** (ethics approval, invitation email, consent form, questionnaire). Raw data and study materials are also available at https://github.com/SarahAlzahranitcd/Thesis-Materials/tree/main/Study-2-Usability

[7] https://virtualtreasury.ie

5.4.4 Evaluation instruments

Both quantitative and qualitative data were collected during the experiment.

System usability and satisfaction were measured using a 14-item instrument combining elements from two established questionnaires: the **Post-Study System Usability Questionnaire (PSSUQ)** [76] and **the System Usability Scale (SUS)** [77]. PSSUQ was selected as the primary basis because it provides diagnostic subscale scores across three dimensions (System Usefulness, Information Quality, and Interface Quality) which allow usability issues to be localised rather than producing a single overall score as SUS does.

Items from the SUS were incorporated to strengthen coverage of ease of use and learnability, which are directly relevant to evaluating a tool intended for non-expert users. The resulting 14-item instrument uses a 1-7 Likert scale where lower scores indicate higher perceived usability.

Four items are negatively phrased: Q8 (unnecessarily complex), Q9 (required extensive prior learning), Q11 (too much inconsistency), and Q13 (challenging to use). These items were included following standard psychometric practice to guard against acquiescence bias, the tendency for participants to agree with any statement regardless of content. Prior to analysis, these items were reverse-coded using the formula RC = 8 - original score, so that all items are oriented in the same direction with lower values consistently indicating better usability.

Internal consistency of the instrument was assessed using Cronbach's alpha [74], with a threshold of $\alpha \geq 0.70$ indicating acceptable reliability.

Additional quantitative measures included **task completion time**, derived from server-side timestamps embedded in each participant's annotation output file, and **metadata representation preference** captured through a post-task survey question.

Qualitative data consisted of open-ended responses to a final survey question. These were analysed using **thematic analysis** [78] to identify recurring patterns across participant feedback.

5.4.5 Experiment Structure

The experiment was structured as follows:

1. **Informed consent:** At the beginning of the experiment, participants were provided with an explanation of the study purpose and were required to confirm consent before proceeding.
2. **Scenario allocation:** Each participant was randomly assigned one of three scenarios representing a different declarative mapping type by the system. Each scenario included a mapping file and a contextual description of the project that produced it. Table 5.9 summarises the three scenarios.
3. **Metadata annotation task:** After reading the scenario, participants downloaded the mapping file, uploaded it to MetaSEMAP, and completed the metadata fields based on the scenario. The fields were aligned with the lifecycle phases defined in LMMD.
4. **Review of generated metadata:** After completing the annotation form, participants reviewed the generated metadata output before clicking on the submit button. The result page displayed the annotated mapping in both RDF-star and Named Graph representations. Participants could then view or export the files in Turtle format.
5. **Questionnaire:** After reviewing the generated output, participants completed the usability instrument consisting of 14 questions on 1-7 Likert scale, along with a question on representation preference.
6. **Open-ended feedback:** At the end of the experiment, participants were given the opportunity to submit additional comments on their experience, the clarity of the metadata fields, and suggestions for improvement.

Table 5.9 Summary of the three scenarios

Mapping Type	Scenario Summary
Uplift Mapping	Convert county information from historical data into RDF triples. The file data_county.csv, containing county identifiers, names, and geographic details, was mapped using RML. The resulting mapping enriches the VRTI Knowledge Graph with geospatial and semantic county records.
Ontology Alignment	Align person entities between the VRTI Knowledge Graph and external datasets. Features such as relationships, affiliations, and roles were matched using the file person_alignment.rdf, with the goal of enabling integrated access across datasets.
Interlinking	Connect Irish historical figures in VRTI to corresponding Wikidata entries using owl:sameAs relationships. The file interlinking.rdf was used to establish cross-referenced biographical and contextual data.

5.4.6 Participants and Recruitment

Participants were MSc students enrolled in a Knowledge and Data Engineering course at Trinity College Dublin. Participation was voluntary, anonymous, and completed online. No personally identifying information was collected. Participants completed the experiment asynchronously using the publicly available MetaSEMAP website and were informed that task duration was recorded from the start of the annotation process until submission of the annotation form.

A total of 50 students participated in the experiment. They had introductory knowledge of declarative mappings, including uplift mappings, ontology alignment, and interlinking, but varied in their level of technical experience. This participant group was considered appropriate for this evaluation for two reasons. First, MetaSEMAP is designed to facilitate annotation by those who may lack expertise in mapping, therefore assessing its usability with non-expert users aligns with the tool's target audience. Second, MSc students in a knowledge and data engineering course represent practitioners who are likely to encounter declarative mappings in professional or research settings, making their engagement with the tool meaningful and valid. Out of 50 participants, 46 completed the post-task

questionnaire and are incorporated into the usability analysis. The remaining four completed the annotating task but failed to submit the survey.

5.4.7 Results

The following subsections present the findings of the usability evaluation, organised as follows: system usability and user satisfaction; task completion time; metadata representation preference; metadata annotation observations; and qualitative feedback. All analyses are based on the 46 participants who completed both the annotation task and the post-task questionnaire. The instrument demonstrated excellent internal consistency (Cronbach's $\alpha = 0.899$), confirming that the 14 items reliably measure a coherent construct.

A. System Usability and User Satisfaction

System usability was assessed using the 14-item combined PSSUQ and SUS instrument on a 1-7 scale where lower scores indicate higher perceived usability. Table 5.10 presents the PSSUQ scores per participant.

Table 5.10 PSSUQ scores per participant
(n = 46). SysUse = System Usefulness; InfoQual = Information Quality; IntQual = Interface Quality

Participant	SysUse	InfoQual	IntQual	Overall	Scenario
P01	1.25	2.25	1.00	1.43	Uplift Mapping
P02	2.50	4.25	1.75	2.93	Interlinking
P03	2.50	2.50	1.50	2.00	Uplift Mapping
P04	1.50	1.25	2.50	2.14	Interlinking
P05	1.75	2.00	1.75	1.79	Uplift Mapping
P06	2.75	4.25	2.50	3.14	Uplift Mapping
P07	3.75	4.25	3.75	4.07	Ontology Alignment
P08	1.00	1.25	1.00	1.07	Ontology Alignment
P09	2.25	3.25	3.50	3.00	Uplift Mapping
P10	1.25	2.50	1.00	1.50	Ontology Alignment
P11	1.00	2.50	1.00	1.43	Uplift Mapping
P12	2.25	2.00	2.00	2.07	Uplift Mapping
P13	4.50	4.50	3.75	4.14	Ontology Alignment
P14	2.00	2.00	2.00	2.00	Interlinking
P15	2.00	3.75	3.75	3.14	Ontology Alignment
P16	1.00	2.75	2.50	1.93	Uplift Mapping
P17	4.75	3.75	4.50	4.43	Interlinking
P18	4.75	5.25	4.75	5.00	Uplift Mapping
P19	2.50	2.25	2.25	2.21	Uplift Mapping
P20	2.50	3.75	4.00	3.21	Ontology Alignment
P21	1.00	1.25	1.00	1.07	Ontology Alignment
P22	1.50	1.75	2.75	1.86	Ontology Alignment
P23	2.25	3.25	2.50	2.64	Ontology Alignment
P24	4.25	3.75	4.25	4.00	Interlinking
P25	2.25	3.00	3.25	2.71	Interlinking
P26	3.75	5.00	4.25	4.07	Interlinking
P27	1.00	4.00	4.00	2.71	Uplift Mapping
P28	1.50	3.25	2.50	2.57	Uplift Mapping
P29	1.75	1.75	1.75	1.71	Interlinking
P30	2.00	1.50	2.00	1.86	Ontology Alignment
P31	1.25	2.50	2.00	2.00	Ontology Alignment
P32	1.50	2.00	1.75	1.86	Ontology Alignment
P33	1.00	2.00	1.50	1.50	Ontology Alignment
P34	2.50	3.00	2.25	2.57	Ontology Alignment
P35	1.50	2.75	1.75	2.00	Uplift Mapping
P36	1.50	2.50	1.50	1.86	Uplift Mapping
P37	1.50	1.50	1.50	1.50	Interlinking
P38	2.00	3.00	3.50	2.79	Uplift Mapping
P39	2.50	2.75	4.00	3.07	Uplift Mapping
P40	4.75	4.00	4.00	4.21	Interlinking
P41	2.50	3.00	2.25	2.57	Ontology Alignment
P42	2.50	3.00	2.00	2.43	Ontology Alignment
P43	1.25	2.50	2.25	2.00	Uplift Mapping
P44	1.25	2.50	2.00	1.86	Uplift Mapping
P45	1.50	1.50	2.25	1.71	Interlinking
P46	1.00	4.00	4.00	2.71	Uplift Mapping

Mean	2.15	2.85	2.55	2.49	n = 46
SD	1.09	1.03	1.09	0.95	Sufficiency threshold: < 3.5

Table 5.11 summarises the mean and standard deviation for each subscale across all 46 participants.

Table 5.11 PSSUQ subscale means and standard deviations (n = 46).

Category	Questions Used	Mean	SD	Threshold Met (< 3.5)
System Usefulness	Q4, Q5, Q6, Q7	2.15	1.09	Yes
Information Quality	Q1, Q2, Q8, Q9	2.85	1.03	Yes
Interface Quality	Q3, Q10, Q11, Q13	2.55	1.09	Yes
Overall Usability	All of the above + Q12, Q14	2.49	0.95	Yes

The overall usability score was 2.49 (SD = 0.95), well below the 3.5 sufficiency threshold. System Usefulness received the strongest score of 2.15 (SD = 1.09), indicating that participants found the tool effective and capable of supporting the annotation task. Interface Quality was rated 2.55 (SD = 1.09), reflecting a consistent and navigable layout. Information Quality received a comparatively higher score of 2.85 (SD = 1.03). The weakest individual item was Q9 (prior learning required, reverse-coded M = 3.67), the only item to exceed the 3.5 threshold, indicating that some participants felt they needed to learn about the metadata concepts before completing the task effectively. The strongest items were Q3 (consistent presentation, M = 1.89), Q6 (easy to learn, M = 1.98), and Q14 (overall satisfied, M = 2.11). Table 5.12 presents the mean for each of the 14 items.

Table 5.12 Per-item means for the 14-item PSSUQ/SUS instrument (n = 46).

Item	Description	Direction	Mean	SD
Q1	The tool provided error messages that clearly guided me on how to resolve issues	Positive	2.65	1.42
Q2	Whenever I made a mistake, I could recover easily and quickly	Positive	2.35	1.22
Q3	The tool is consistent in the way it presents information and choices	Positive	1.89	0.99
Q4	The tool provides all the functions I need for annotating mapping files	Positive	2.39	1.24
Q5	The tool is easy to use for annotating mapping files	Positive	2.13	1.44
Q6	It is easy to learn how to use the tool	Positive	1.98	1.16
Q7	Most people would learn to use the tool very quickly	Positive	2.11	1.12
Q8	I find the tool unnecessarily complex (RC)	Negative	2.74	1.69
Q9	I needed to learn a lot of things before I could effectively use the tool (RC)	Negative	3.67	1.92
Q10	The various functions in the tool are well-integrated	Positive	2.59	1.33
Q11	There was too much inconsistency in the tool (RC)	Negative	2.78	1.71
Q12	I felt very confident using the tool	Positive	2.50	1.36
Q13	I found the tool challenging to use (RC)	Negative	2.96	1.84
Q14	Overall, I am satisfied with my experience using the tool	Positive	2.11	1.39

B. Task Completion Time

Completion times were derived from server-side timestamps embedded in each participant's annotation output file, recording elapsed time from the start of the annotation session to file submission.

The initial mean completion time across the 46 survey participants was 46.3 minutes (SD = 101.9). Six submissions completed in under five minutes were excluded as likely incomplete or interrupted (participants 017Y, 815M, 0021, 054T, 010X, 361R), and three submissions exceeding 200 minutes were excluded as probable interruptions (participants 771E, 778W, 880B). After removing these nine outliers, the valid sample comprised 37 participants.

Table 5.13 presents the adjusted completion times by scenario. Uplift Mapping participants completed the task in a mean of 21.6 minutes (SD = 19.9, Median = 14.2, n = 16). Ontology Alignment participants took a mean of 18.8 minutes (SD = 11.4, Median = 15.4, n = 12). Interlinking participants took substantially longer on average (M = 39.9 min, SD = 42.9, Median = 21.9, n = 9), though the high standard deviation and median of 21.9 minutes indicate that one participant (139.6 minutes) drove the mean upward. The longer times for Interlinking are consistent with the greater conceptual complexity of that scenario, which requires reasoning about entity-level correspondences across datasets rather than the more procedural field-filling of the uplift and alignment tasks.

Table 5.13 Task completion times by scenario

Scenario	n	Outliers Removed	Adjusted Mean (min)	SD	Median (min)	Range (min)
Uplift Mapping	16	1	21.6	19.9	14.2	5.1 – 82.6
Ontology Alignment	12	2	18.8	11.4	15.4	6.7 – 49.0
Interlinking	9	0	39.9	42.9	21.9	5.1 – 139.6
All participants	37	9	25.1	26.2	15.9	5.1 – 139.6

The results provide evidence that **Hypothesis 2** is supported. All 46 participants who initiated the experiment completed the annotation task across the three mapping types, demonstrating that MetaSEMAP supports task completion regardless of mapping type.

C. Preferred Metadata Representation

After completing the annotation task, participants reviewed the generated output in both RDF-star and Named Graph representations and indicated which they found easier to read and interpret. 34 out of the 46 participants (73.9%) expressed a preference for Named Graph and 12 (26.1%) preferred RDF-star. The preference for Named Graph was consistent across all three scenario groups: 74% for Uplift Mapping, 81% for Ontology Alignment, and 64% for Interlinking.

The Named Graph representation produces metadata as a separate file from the mapping specification. RDF-star embeds metadata annotations inline within the mapping file using the <<...>> notation. The preference for Named Graph suggests that participants may have found the separation of metadata from mapping content easier to read and interpret. This finding is discussed further in Section 5.4.8.

The results provide evidence that **Hypothesis 3** is supported. A clear preference for Named Graph was observed across all three scenario groups, with Named Graph preferred by 73.9% of participants overall.

D. Metadata Annotation Observations

The submitted annotation files were reviewed by the author of this thesis to examine how participants completed the metadata fields across the three mapping types and to assess the feasibility of applying the LMMD metadata model to each type.

Annotation feasibility by mapping type

MetaSEMAP supports annotation of all three declarative mapping types. The mechanism for attaching metadata differs by type due to structural differences between mapping formats.

For **uplift mappings** (RML Turtle files), the RDF-star annotation subject is the TriplesMap entity declaration. Metadata is attached to the triple asserting the mapping's primary resource as a *rr:TriplesMap*. For the Named Graph output, the same metadata fields are written as properties of a placeholder subject in a separate Turtle file.

For **ontology alignment files** (using the Alignment API vocabulary), the annotation subject is the alignment instance (*align:Alignment*). The same structured metadata fields are attached, associating lifecycle metadata with the alignment artefact.

For **interlinking specifications** (SPARQL query files), no native RDF resource exists in the source file. MetaSEMAP addresses this by defining a new resource declared as type *InterlinkingOperation*, which then serves as the annotation subject.

All 46 survey participants successfully produced annotated output for their assigned mapping type, confirming that lifecycle metadata annotation is feasible across all three types within the current MetaSEMAP implementation.

It should be noted that the RDF-star output uses the <<...>> notation at the mapping entity level rather than at the level of individual RDF triples. This represents a project-level metadata attachment approach using RDF-star syntax, rather than triple-level reification. The Named Graph output uses a Turtle serialisation with a generic placeholder subject (metag:subject), as illustrated in Listing 4.3. The preference for Named Graph expressed by participants is consistent with the separation of metadata from the mapping file that the Named Graph output provides, and suggests that keeping metadata as a standalone artefact supports

readability. Both representations are acknowledged as initial implementation choices made to enable rapid deployment and evaluation of the tool. MetaSEMAP v2 addresses the Named Graph limitation by producing conformant TriG output with a proper named graph URI, which formally separates the metadata graph from the default graph and enables graph-level querying and provenance tracking.

Field completion patterns

Structured fields including *start* and *end dates, mapping URI, version number, publisher name,* and *testing type* were generally completed with high consistency across all participants.

The *mapping purpose* field showed the greatest variability across scenarios. Uplift Mapping participants generally provided adequate or detailed descriptions (86% of submissions), as did Ontology Alignment participants (94%). Interlinking participants showed substantially more variability, with 63% providing generic or minimal descriptions. Phrases such as "Enhance the historical dataset" or the single word "Enrichment" appeared repeatedly.

This pattern likely reflects the greater conceptual distance between the SPARQL query file and its broader project purpose, which was less immediately apparent to participants than the structured field-to-file correspondence in uplift and alignment scenarios. Six of the 46 participants selected an incorrect mapping method, suggesting that the distinction between manual and automatic approaches requires clearer in-tool guidance.

E. Qualitative Feedback

31 participants out of the 46 provided qualitative comments. Responses were coded inductively using thematic analysis.

Table 5.14 presents the six themes identified. Ease of use and simplicity was the most frequently referenced theme (15 references), with participants describing the tool as accessible and non-technical. Field guidance and example values (8 references) and interface layout improvements (6 references) were the next most common, consistent with the Information Quality and Interface Quality subscale results. Requests for additional features (5 references) included reusable templates, collaboration support, and enhanced consistency checks. Error message clarity (3 references) echoed the weaker score on Q1 (error messages, M = 2.65). Scenario visibility (2 references) reflected a concern about maintaining context during annotation.

Table 5.14 Themes identified from qualitative feedback (n = 31).

Theme	References	Representative Comment
Ease of use and simplicity	15	"Everyone can use this tool as it is very easy to use and the user does not require any technical knowledge."
Field guidance and example values	8	"Add explanatory descriptions for each input field like the question mark for sections, but for each individual field."
Interface layout improvements	6	"Increase the size of text input boxes; add auto-wrapping for long strings."
Additional features	5	"Add predefined templates, collaboration support, and enhanced consistency checks."
Error message clarity	3	"Add a URI validator for required fields; provide cleaner error handling messages."
Scenario visibility	2	"Please have the scenario on the same page as the annotation form."

5.4.8 Discussion

Hypothesis 1 (Usability). Hypothesis 1 is supported. The overall PSSUQ score of 2.49 (SD = 0.95, $\alpha = 0.899$) falls below the 3.5 sufficiency threshold, and this result holds across all three scenario groups. System Usefulness (M = 2.15) and Interface Quality (M = 2.55) received the strongest scores, confirming that participants found the tool effective and its layout consistent. Information Quality (M = 2.85) was comparatively weaker, driven primarily by Q9 (prior learning required, RC M = 3.67) and Q1 (error message guidance, M = 2.65). These two items identify specific areas for improvement rather than a systemic usability problem. The higher variability observed in the Interlinking group (SD = 1.13) compared to Uplift Mapping (SD = 0.82) and Ontology Alignment (SD = 0.94) is consistent with the greater conceptual complexity of that scenario which requires attention in future iterations.

Hypothesis 2 (Task effectiveness). Hypothesis 2 is supported. All 46 participants completed the annotation task across all three mapping types. The adjusted completion times confirm meaningful engagement, with median times of 14.2 minutes for Uplift Mapping, 15.4 minutes for Ontology Alignment, and 21.9 minutes for Interlinking. The longer times for Interlinking participants reflect the conceptual demands of reasoning about entity-level correspondences across datasets, a task that is inherently more abstract than the field-to-file correspondence of the uplift and alignment scenarios.

Hypothesis 3 (Representation preference). Hypothesis 3 is supported. A clear preference for Named Graph was observed across all three scenario groups (73.9% overall). The Named Graph representation produces metadata as a standalone file, physically separated from the mapping specification. The RDF-star representation embeds metadata inline within the mapping file. The preference for Named Graph is consistent with the principle that metadata should be maintained as an independently readable and queryable artefact, which is a core design goal of LMMD. However, the comparison conditions in this study were not fully

controlled. Prefix conventions differed between the two outputs, and the physical separation of Named Graph metadata may have contributed to its perceived readability independently of the format itself. A more controlled comparison is needed before drawing strong conclusions about the relative readability of the two formats. However, the refined version of the tool allows users to export metadata in named graph.

Broader observations. The annotation quality review revealed that structured fields were completed with high consistency, confirming that the interface effectively supports the required metadata fields. The variability in the mapping purpose field, particularly among Interlinking participants where 63% provided generic descriptions, indicates that certain metadata concepts require additional in-tool guidance such as example values or field-level prompts. The six method selection errors confirm that the manual/automatic distinction needs clearer definition. These findings directly informed refinements to MetaSEMAP v2, which introduced additional in-tool guidance including example values for the mapping purpose field, clearer labelling of the manual/automatic method distinction, and field-level tooltips to reduce annotation ambiguity.

5.4.9 Experiment Summary

Study 2 evaluated the usability of MetaSEMAP as an implementation of the Lifecycle-Based Metadata Model for Declarative Mappings (LMMD). The evidence gathered suggests that all three hypotheses were supported. The overall PSSUQ score of 2.49 (SD = 0.95, $\alpha = 0.899$) confirmed sufficient usability across all three mapping types. Structured metadata fields were completed with high consistency, while the mapping purpose field revealed areas where additional guidance would improve precision. A clear preference was observed for the Named Graph representation over RDF-star (73.9% versus 26.1%), consistent with the separation of metadata from the mapping file that Named Graph provides.

Limitations of this study include the use of a student participant sample and controlled experimental scenarios, which may not fully reflect complex real-world mapping projects. The comparison between RDF-star and Named Graph representations was not fully controlled in terms of file structure and prefix conventions, which limits the strength of conclusions that can be drawn from the representation preference result. These limitations do not undermine the primary finding that the lifecycle-based metadata annotation is feasible and usable across all three mapping types but should be addressed in future iterations.

While Study 1 established the perceived relevance of lifecycle-based metadata at a conceptual level, Study 2 demonstrated its usability in practice through tool-based implementation. This provides a foundation for the subsequent ontology-level validation presented in Study 3.

5.5 Study 3: Ontology Evaluation

This section presents an ontology evaluation study of the Mapping Metadata Vocabulary (MMV), the formal ontology implementing the Lifecycle-Based Metadata Model for Declarative Mappings (LMMD). The focus of this experiment was on evaluating whether MMV adequately represents lifecycle-based metadata for declarative mapping projects and whether its conceptual structure, modelling choices, and vocabulary reuse are clear and appropriate from an ontology engineering perspective. The section is structured into the following subsections: Motivation; Hypotheses; Methodology; Evaluation Instruments; Participants and Recruitment; Results; Discussion; and Experiment Summary.

5.5.1 Motivation

While Study 1 examined the perceived relevance of lifecycle-based metadata fields and Study 2 evaluated their usability through the MetaSEMAP tool, neither study

directly assessed the formal ontology representation of the model. The Mapping Metadata Vocabulary (MMV) represents the formalisation of the Lifecycle-Based Metadata Model for Declarative Mappings (LMMD) as an OWL ontology, explicitly modelling mapping projects, lifecycle activities, and mapping artefacts.

Since MMV introduces a structured, lifecycle-oriented abstraction and reuses existing vocabularies such as PROV-O, DQV, and MQV, it was necessary to evaluate whether the ontology is conceptually clear, adequately structured, and aligned with its intended requirements. In particular, the ontology must correctly represent lifecycle phases, support reuse-oriented documentation, and maintain appropriate abstraction across different mapping types.

The motivation of Study 3 was therefore to evaluate MMV from an ontology perspective, focusing on conceptual clarity, lifecycle coverage, modelling structure, and vocabulary reuse. In addition to the structured expert evaluation, Study 3 provides an opportunity to assess whether MMV correctly answers the competency questions defined during its design (Section 4.5.3). The SPARQL query templates provided in Appendix B demonstrate that each competency question can be answered using MMV properties, providing formal evidence of ontology coverage.

The ontology formalisation of LMMD as MMV was presented at the SDI Workshop at IEEE ICSC 2026 [83], where an initial description of the ontology design and a planned expert evaluation were outlined. Study 3 reports the full evaluation conducted following that publication.

5.5.2 Hypothesis

The following hypotheses were defined for this experiment:

- **Hypothesis 1 (Conceptual clarity):** Experts will report MMV classes and properties are clear and understandable.

- **Hypothesis 2 (Lifecycle coverage):** Experts will report MMV adequately captures metadata required to document mapping projects across the lifecycle phases of analysis, design, development, testing, and maintenance.
- **Hypothesis 3 (Modelling quality):** Experts will report MMV follows the appropriate ontology engineering practices, including structured lifecycle modelling and reuse of established vocabularies.

These hypotheses reflect the central design claims of MMV as presented in Chapter 4.

5.5.3 Methodology

To test the hypotheses, a structured expert-based evaluation of the MMV ontology was conducted to assess conceptual adequacy, modelling decisions, and alignment with design goals. The evaluation approach adopted in this study follows a phased structure combining automated validation tools, expert review, and qualitative feedback, consistent with the phased ontology evaluation methodology proposed by Randles et al [79].

The experiment was carried out individually and anonymously through an online questionnaire. Participants were provided with access to: the MMV ontology description, an illustrative Turtle instantiation example, and links to the ontology/model specification and supporting documentation before completing the evaluation questionnaire.

Both quantitative and qualitative data were collected. Quantitative data consisted of Likert-scale ratings assessing conceptual clarity, lifecycle coverage, structural organisation, and vocabulary reuse. Qualitative data consisted of open-ended responses allowing participants to suggest improvements, identify ambiguities, and comment on the modelling structure.

In addition to the core ontology evaluation, the questionnaire included an exploratory section investigating how **mapping purpose** should be represented in the ontology. This section examined whether *purpose* should be captured using structured categories rather than free-text descriptions, as part of improving reuse-oriented metadata.

Ethical approval was received from the Research Ethics Committee at Trinity College Dublin. The study materials are provided in Appendix F (ethics approval, invitation email, consent form, OOPS! and FOOPS! screenshots of the results and expert evaluation questionnaire). Raw data and study materials are also available at https://github.com/SarahAlzahranitcd/Thesis-Materials/tree/main/Study-3-Ontology-Evaluation

5.5.4 Evaluation Instruments

The ontology evaluation used two complementary instruments.

- **Automated ontology validation** tools were used to assess syntactic correctness, consistency, and modelling quality. Two tools were applied: OOPS! (Ontology Pitfall Scanner) [80] to detect common ontology pitfalls, and FOOPS! (FAIR Ontology OPeration and Suitability) [81] to assess conformance to FAIR principles for ontologies.
- **An expert review questionnaire** was used to assess conceptual clarity, lifecycle coverage, modelling structure, and vocabulary reuse. The questionnaire used a 1-5 Likert scale, where 1 = Strongly Agree and 5 = Strongly Disagree, and included open-ended questions for additional feedback.

Participants were provided with the MMV ontology documentation, a summary of the lifecycle phases, and an example instantiation showing how MMV represents a mapping project. In addition, a purpose modelling section was included, asking participants to evaluate whether mapping purpose should be represented through a

standalone controlled vocabulary consisting of structured categories (e.g., project goal, domain, intended use) and to suggest additional categories where appropriate.

5.5.5 Experiment Structure

The experiment was structured as follows:

1. **Informed consent:** Participants were provided with an information sheet describing the purpose of the study and confirmed their consent before proceeding.
2. **Review of ontology documentation:** Participants were asked to review the MMV ontology documentation and illustrative example prior to answering the questionnaire.
3. **Questionnaire-based evaluation:** Participants completed Likert-scale questions assessing ontology clarity, coverage, and modelling structure.
4. **Open-ended feedback:** Participants were invited to provide additional comments and suggestions for improvement.
5. **Automated validation:** in parallel, MMV was assessed using OOPS! and FOOPS! to detect ontology pitfalls and evaluate conformance to FAIR principles.

5.5.6 Participants and Recruitment

Participants were recruited through targeted outreach to ontology engineers and researchers working in the Semantic Web and knowledge graph communities. The study aimed to involve participants with prior experience in ontology development or Semantic Web technologies. Invitations were distributed via direct contact with researchers. Participation was voluntary and anonymous. A total of 6 participants completed the ontology evaluation. Participants satisfied the following inclusion criteria: experience in ontology development or Semantic Web research, and

familiarity with RDF-based vocabularies. All participants were researchers with relevant expertise

5.5.7 Results

This section presents the findings of the ontology evaluation of MMV, organised as follows: quantitative evaluation scores; expert responses to modification suggestions; expert responses to documentation suggestions; *purpose* vocabulary findings and automated validation results;

A. Quantitative Evaluation

Table 5.15 presents the scores provided by each expert for the six evaluation statements. The scale ranges from 1 (Strongly Agree) to 5 (Strongly Disagree).

Table 5.15 Quantitative scores for the expert evaluation

Statement	E1	E2	E3	E4	E5	E6	Mean	Median
Q1 Captures essential concepts	2	1	4	1	1	3	2.00	1.5
Q2 Coherent structure	3	1	5	1	1	2	2.17	1.5
Q3 Classes/properties clearly defined	1	1	5	1	1	1	1.67	1.0
Q4 Follows ontology design principles	1	1	-	1	2	3	1.60	1.0
Q5 Reuses established vocabularies	1	1	5	1	1	1	1.67	1.0
Q6 Supports reuse decisions	2	1	5	1	1	2	2.00	1.5

Five of the six participants provided consistently positive ratings. The strongest results were observed for Q3 (classes and properties clearly defined, mean = 1.67, median = 1.0) and Q5 (vocabulary reuse, mean = 1.67, median = 1.0), with five of the six participants selecting the most positive response for both statements. Q4 (ontology design principles, mean = 1.60, median = 1.0) received similarly strong scores; one participant did not answer this question. Q1 (essential concepts, mean = 2.00, median = 1.5) and Q6 (reuse support, mean = 2.00, median = 1.5) remained

positive overall. One participant (E3) provided substantially lower ratings across all statements. When this participant is excluded, means fall below 2.0 for all statements, indicating strong agreement among the remaining five participants.

B. Expert Responses (Modifications to Concepts and Properties Q7)

Table 5.16 presents the responses to Q7 from each participant. Q7 is "Do you suggest any modifications to the concepts, properties or modelling decisions ? Please justify."

Table 5.16 Q7 Modifications

Participant	Suggestions	Comments
E1	None	N/A
E2	None	No
E3	None	--
E4	None	--
E5	Minor	"Overall, I think the ontology provides a strong and well structured model for representing Mapping Lifecycle Metadata. It covers the essential concepts and properties needed to model mappings across different stages of the lifecycle. The reuse of well-known ontologies strengthens the design by improving interoperability and alignment with existing standards. However, I noticed that there are a lot of properties with ranges of type string. While I understand that this may be necessary in certain situations, I would recommend reviewing them to ensure that the range is suitable for the property usage. For instance, the hasInputCreator property could have an object range, providing the ability to attach metadata to the creator (affiliation, email) and facilitate deduplication across systems (two people with the same name)"
E6	Yes	"I would avoid using string for the hasConcept properties (e.g. hasPurpose). It is really important to have a concept for "Purpose", "Domain", etc. and then use strings for as their values if there are no vocabularies/taxonomies to populate them from. The definition of hasStakeholder is missing. I would suggest getting a permanent URL for the ontology from w3id and also submit it to LOV for better accessibility and discoverability."

Two participants provided feedback. E5 noted that the overall structure was strong but suggested that several data properties with string ranges, such as *hasInputCreator*, would benefit from object property ranges to support richer metadata and deduplication across systems. E6 recommended that concept-

oriented data properties such as *hasPurpose* and *hasMappingDomain* be changed to object properties linking to named resources rather than string literals, in order to improve semantic precision and enable linking to external vocabularies. This recommendation is consistent with the findings of Study 1, where practitioners identified purpose and domain as the highest-priority metadata fields.

C. Expert Responses (Documentation Suggestions Q8)

Table 5.17 presents the responses to Q8 from each participant. Q8 is "Do you suggest improvements to the documentation (examples, explanations, constraints)? Please justify your response."

Table 5.17 Q8 Documentation

Participant	Comments
E1	N/A
E2	"I was wondering how it could connect to the actual SHACL statements that are executed for validation and testing and maybe even the SHACL validation report as a result of a validation process."
E3	--
E4	"Perhaps some examples in the documentation itself, similar to the PROV ontology"
E5	"The documentation includes clear examples and diagrams that show how the ontology can be applied in real-world scenarios and illustrate how key classes and properties are instantiated in practice and help users understand the models structure. Adding SHACL shape examples to validate the model usage would further strengthen the documentation by enabling validation of correct implementation"
E6	"In section 5, providing a human-readable description of the example might be helpful. I also think it is very useful to include the competency questions from the paper in the html documentation."

Four participants provided feedback to improve the documentation of the ontology across three areas. E2 suggested connecting the testing phase metadata to SHACL validation reports and their outcomes. E4 requested examples embedded directly in the documentation, similar to the PROV-O specification. E6 suggested providing a human-readable description of the Turtle example and including the competency questions from the published paper in the HTML documentation. E5 provided a positive assessment, noting the documentation was clear and well-illustrated with real-world scenarios.

D. Purpose Vocabulary

This section presents the results of the exploratory section investigating how mapping purpose should be represented in an ontology. Five purpose categories were proposed for evaluation: **Project Goal** (e.g., data integration, enrichment), **Application Context or Domain** (e.g., research, operational system, decision support), **Intended Use** (e.g., analysis, publication, data exchange), Target Audience (e.g., domain experts, developers, public users), and **Reuse Expectations** (e.g., exploratory reuse, production reuse, one-off use). Participants were asked five questions: (1) whether mapping purpose should be captured using a controlled vocabulary; (2) which of the five proposed categories they considered useful; (3) how they ranked those categories by importance; (4) whether the vocabulary should be applicable beyond mapping projects to Semantic Web artefacts generally; and (5) how such a vocabulary could be used in practice.
Table 5.18 presents the per-expert responses to the controlled vocabulary and broader applicability questions.

Table 5.18 Purpose Vocabulary response

Question	Strongly Agree	Agree	Neutral	Disagree
Purpose should use controlled vocabulary	2	2	1	1
Vocabulary applicable beyond mapping projects	2	2	1	1

Four out of six participants agreed or strongly agreed that mapping purpose should be explicitly captured using a controlled vocabulary. Five out of six participants selected all five proposed categories as useful. E1 selected three categories: Project Goal, Application Context, and Reuse Expectations. Table 5.19 presents the importance rankings for each purpose category per expert (1 = Most Important, 5 = Least Important).

Table 5.19 Purpose category rankings per expert

Participant	Project Goal	Application context	Intended use	Target Audience	Reuse Expectations
E1	2	3	4	5	1
E2	-	-	-	-	-
E3	5	3	4	1	2
E4	5	1	2	3	4
E5	1	3	2	4	5
E6	1	3	2	4	5

Intended Use was most consistently rated as high priority, ranked first or second by four out of five experts who provided rankings. Project Goal showed the most variation, ranked first by two experts and fifth by two others.

Table 5.20 presents the open-ended expert responses on suggestions for additional purpose categories, or comments on how purpose metadata could support reuse and interpretation of mappings.

Table 5.20 Expert open-ended responses on additional purpose categories

Participant	Comments
E1	--
E2	"I think the controlled vocabulary could be useful for reuse and for bootstrapping developers/ontologists getting started with this model but eventually organisations might want to define their own controlled vocabulary / taxonomy, with their own specific terms, so there should be support and flexibility to do that."
E3	--
E4	"Perhaps links to in-use projects."
E5	"Overall, I think these categories provide a comprehensive list of purposes. Purpose metadata can help users determine whether a mapping is appropriate to reuse for their specific use case."
E6	--

Table 5.21 presents the open-ended expert responses on how a purpose vocabulary could be used in practice.

Table 5.21 Expert open-ended responses on purpose vocabulary usage

Participant	Comments
E1	Purpose Alignment (understand the artifact purpose and domain of knowledge), Catalogue
E2	--
E3	--
E4	Explicitly modelling purpose improves discovery and reuse by enabling users to identify resources aligned with specific use-cases. It also supports clearer interpretation of design intent, reduces misapplication, and strengthens governance by informing lifecycle management, sustainability planning, and strategic decision-making.
E5	Using a controlled vocabulary to represent the purpose of a mapping artefact may reduce flexibility, make it more difficult to capture new purposes and require ongoing maintenance to remain relevant.
E6	Using a general-purpose, high-level vocabulary would bring the common language and interoperability required across different stages of Semantic Web-based projects. However, if the types of purposes are too different or conflicting to be included in one vocabulary, I would prefer a modular approach wherein there are multiple vocabularies/taxonomies of purposes but all are specialisation of the main ontology (e.g. extending the Purpose concept).

E. Automated Validation

MMV was assessed using OOPS! and FOOPS! in parallel with the expert review. OOPS! identified no critical pitfalls in MMV-native classes or properties. The three critical flags detected (P19) relate to multiple domain declarations on properties inherited from PROV-O (prov:hadActivity, prov:hadRole, prov:wasRevisionOf) and reflect a known characteristic of that vocabulary rather than MMV design errors. FOOPS! confirmed full conformance with FAIR principles, achieving 100% across all evaluated dimensions. Screenshots of both results are provided in Appendix F.

5.5.8 Discussion

The expert evaluation provides encouraging evidence that MMV is conceptually clear, structurally coherent, and aligned with its intended design goals. The strongest results were observed for the clarity of classes and properties and for the

reuse of established vocabularies, which directly supports the claim that MMV was designed according to recognised ontology engineering principles. The positive scores for lifecycle coverage also suggest that experts considered the ontology capable of representing the key concepts required to document mapping projects across analysis, design, development, testing, and maintenance.

At the same time, the expert feedback points to several useful areas for refinement. One recurring issue concerns the use of string valued properties for concepts that may benefit from explicit modelling. In particular, suggestions to represent purpose, domain, and creator-related information through object properties rather than plain strings are important, as they would improve semantic precision, interoperability, and the possibility of linking to external vocabularies.

The results from the purpose modelling section reinforce this point. Experts clearly indicated that mapping purpose is essential for reuse and should be represented in a more structured and semantically explicit way. This suggests that purpose-related metadata should evolve beyond descriptive text toward a more formal representation aligned with controlled vocabularies or domain-specific ontologies.

The documentation related feedback is also valuable. Although the ontology was rated positively overall, several experts requested stronger examples, more human-readable explanations, and explicit SHACL-based validation examples. This suggests that the ontology itself is not the main issue; rather, additional guidance is needed to support correct instantiation and usage.

The results also indicate that MMV is perceived as potentially useful for supporting reuse decisions, although this dimension showed slightly more variation. This is expected, as reuse decision-making depends not only on the ontology structure but also on the completeness and quality of metadata instances.

The purpose vocabulary section revealed an important gap in the current Semantic Web landscape. During the development of LMMD and MMV, it was observed that no established controlled vocabulary exists for representing the purpose of

semantic artefacts. The expert responses confirmed this gap, four out of six participants agreed that purpose should be explicitly modelled, and five out of six considered the proposed categories useful. This finding extends beyond mapping artefacts. Purpose is a critical interpretive element for any semantic artefact, including ontologies, datasets, mappings, and workflows, yet it is consistently treated as free text or omitted entirely. The research community would benefit from a standardised vocabulary for representing purpose across all types of semantic artefacts, enabling better discovery, governance, reuse, comparison and FAIR compliance. This is identified as a direction for future work arising from this thesis.

Overall, the findings support the three hypotheses of this study. MMV was generally viewed as clear and understandable, was seen as providing adequate lifecycle coverage, and was judged to follow appropriate ontology design practices.

5.5.9 Experiment Summary

Study 3 evaluated MMV as the formal ontology implementing LMMD. Expert evaluation confirmed that the ontology is conceptually clear, structurally coherent, and aligned with its design goals, with the strongest results observed for class and property clarity and vocabulary reuse. Automated validation using OOPS! confirmed no critical pitfalls in MMV-native properties. FOOPS! confirmed full conformance with FAIR principles, achieving 100%.

The qualitative feedback informed several refinements incorporated into MMV v0.2: hasValidationReport was added as an object property on TestingActivity to connect testing metadata to SHACL validation reports, and the ontology documentation was updated with richer examples and competency questions as requested by E4 and E6. The suggestion to replace string-valued properties with a formally structured purpose vocabulary is identified as future direction, as no suitable controlled vocabulary currently exists in the Semantic Web community

for this purpose. The revised ontology is available at https://sarahalzahranitcd.github.io/MMV-V02/

The purpose vocabulary section revealed a broader gap in the Semantic Web community. No standardised controlled vocabulary exists for representing the purpose of semantic artefacts, a future research direction identified by this thesis.

Study 3 provides ontology-level evidence that LMMD can be represented in a conceptually strong and formally structured way. This provides a foundation for Study 4, which investigates whether lifecycle-based metadata can support mapping reuse decisions in practice.

5.6 Study 4: Mapping Reuse Experiment Based on Metadata

This section presents a scenario-based expert study investigating whether lifecycle-based metadata supports reuse decision-making for declarative mapping artefacts. The study compares two metadata representations a FAIR-oriented artefact-level representation and the MMV lifecycle-based metadata model across the three mapping types. Participants were asked to assess each artefact for reusability based on the provided metadata descriptions. The section is structured into the following subsections: Motivation; Hypotheses; Methodology; Evaluation Instruments; Experiment Structure; Participants and Recruitment; Results; Discussion; and Experiment Summary.

5.6.1 Motivation

Declarative mappings are often shared as standalone technical artefacts with limited contextual documentation, making it difficult for users to understand their purpose or assess their suitability for reuse. While previous studies demonstrated that lifecycle-based metadata is perceived as relevant by practitioners and can be

applied in practice through MetaSEMAP, it remained necessary to evaluate whether it actually supports reuse decision-making in practice.

The motivation of this study was therefore to investigate whether lifecycle-based metadata improves the ability of experts to make informed reuse decisions when comparing existing mapping artefacts across different mapping types.

5.6.2 Hypothesis

The following hypotheses were defined for this experiment:

1. **Hypothesis 1 (Reuse decision support):** Participants will be better able to judge whether a mapping artefact is reusable when it is described using lifecycle-based metadata.
2. **Hypothesis 2 (Representation preference):** Participants will prefer the lifecycle-based metadata representation over the concise FAIR-oriented metadata representation when making reuse decisions.
3. **Hypothesis 3 (Decision confidence):** Lifecycle-based metadata will provide participants with greater confidence and clearer justification for their reuse decisions.

These hypotheses reflect the central claim of LMMD that structured lifecycle-based metadata improves the transparency and interpretability of declarative mapping artefacts for reuse.

5.6.3 Methodology

To test these hypotheses, a scenario-based expert study was conducted. Participants were presented with three reuse scenarios, each representing a different type of declarative mapping artefact: an interlinking rule, an RML

mapping, and an ontology alignment. In each scenario, the same artefact was described using two different metadata representations:

- Representation A: a FAIR-oriented metadata representation based on FAIR-IMPACT-style artefact-level metadata
- Representation B: the MMV lifecycle-based metadata model, which provides metadata across the mapping lifecycle

Representation A was selected as the comparison baseline because it reflects current FAIR-oriented mapping documentation practice. FAIR-IMPACT [40] recommended extending SSSOM [49] with complementary artefact-level metadata to support findability and reuse across infrastructures, as reviewed in Section 3.1.3. SSSOM itself focuses on correspondence-level and mapping-set-level fields and does not organise metadata across lifecycle phases. Representation A was therefore constructed to reflect the artefact-level metadata profile recommended by FAIR-IMPACT, representing what practitioners would currently be expected to provide. By comparing it against Representation B, this study evaluates whether lifecycle-based metadata improves reuse decision-making beyond what current FAIR-oriented practice already provides.

Participants were asked to review the mapping artefact and its two metadata descriptions, and then decide whether the artefact could be reused as a starting point in a new project. Importantly, they were instructed not to judge the full technical correctness of the artefact itself, but rather whether the metadata description gave them enough information to make an initial informed reuse decision.

The study collected both quantitative and qualitative data. Quantitative responses captured reuse judgements, representation preference, understandability, confidence, and whether the two metadata models were seen as complementary. Qualitative responses captured participants' explanations of what information was useful, what was missing, and any additional comments.

Both metadata representations were presented to participants as structured text tables, with field values populated based on information available from the source artefacts used in each scenario. Where information was not available or not applicable, fields were marked as 'Not provided'. This was an intentional design choice: the varying levels of completeness across scenarios allowed the study to assess whether metadata completeness affects reuse judgement and decision confidence in practice. The full scenario descriptions, mapping artefact excerpts, and both metadata representations for each scenario are provided in Appendix G.

The experiment was carried out online and asynchronously. Ethical approval was received from the Research Ethics Committee at Trinity College Dublin. The study materials are provided in Appendix G (ethics approval, invitation email, consent form, scenario descriptions and questionnaire). Raw data and study materials are also available at https://github.com/SarahAlzahranitcd/Thesis-Materials/tree/main/Study-4-Reuse-Experiment

5.6.4 Evaluation Instruments

The study used a structured questionnaire embedded within each scenario page. The questionnaire combined fixed-choice and open-ended items.

The main evaluation instruments included:

- **Reuse decision question**, asking whether the artefact would be considered reusable as a starting point
- **Representation comparison questions**, asking which metadata representation better supported the reuse decision, was easier to understand, and provided more confidence.
- **Complementarity question**, asking whether the two metadata models were perceived as complementary .
- **Open-ended questions** asking what information was most useful, what was missing, and any additional comments.

These instruments were designed to assess not only reuse judgement itself, but also how the metadata representations influenced participants' reasoning and interpretation.

5.6.5 Experiment Structure

The experiment was structured as follows:

1. **Study information and consent:** Participants first reviewed an information and consent page explaining the purpose of the study, the meaning of mapping artefacts, and the difference between the two metadata representations.
2. **Scenario review:** Participants completed three scenarios, each corresponding to a different mapping type:
    - Scenario 1: Interlinking rule
    - Scenario 2: RML uplift mapping
    - Scenario 3: Ontology alignment.

    The three scenarios were based on real-world mapping artefacts. Scenario 1 used a Silk interlinking specification linking DBpedia and GeoNames city entities example available at [31]. Scenario 2 used a YARRRML/RML mapping from the ERA knowledge graph project, available via the RML.io[8] use cases page. Scenario 3 used the LogMap output from the OAEI 2024 Anatomy track[9].
3. **Artefact and metadata comparison:** In each scenario, participants reviewed:
    - an excerpt of the mapping artefact itself.
    - Representation A (FAIR-oriented metadata)
    - Representation B (MMV lifecycle-based metadata)

[8] https://rml.io/usage/
[9] https://oaei.ontologymatching.org/2024/results/anatomy/index.html

4. **Reuse questionnaire**: After reviewing each scenario, participants answered questions about:
    - Whether they would consider the artefact reusable.
    - Which representation better supported the decision.
    - Which was easier to understand.
    - Which gave more confidence.
    - Whether the two metadata models complemented each other.
5. **Open-ended feedback:** Participants were also asked what information was most useful, what was missing, and whether they had any additional comments.
6. **Submission:** Responses were submitted anonymously after completion of the final scenario.

5.6.6 Participants and Recruitment

Participants in this study were recruited through direct outreach to researchers and practitioners with expertise in declarative mappings, Semantic Web technologies, and related areas. Invitations were also distributed via the semantic-web@w3.org mailing list. Participation was voluntary and anonymous. A total of 10 participants completed the experiment. Their responses provided both comparative judgements across the two metadata representations and qualitative insights into what metadata is considered most useful for reuse decisions.

5.6.7 Results

A total of 10 participants completed all three scenarios. Results are presented question by question below. Tables 5.22, 5.23, and 5.24 present the individual responses for each participant across the three scenarios that answer all questions from question 1 to 5.

Table 5.22 Scenario 1(Interlinking Rule) - Participant responses

P	Rep A Reuse	Rep B Reuse	Q2 Better	Q3 Easier	Q4 Confidence	Q5 Complementary
P1	Yes	Yes	B	Equal	B	No overlap
P2	Yes	Yes	Equal	Equal	Equal	Somewhat
P3	Yes	No	A	A	A	Somewhat
P4	Reservations	Yes	B	B	B	Somewhat
P5	Yes	Reservations	A	B	A	Somewhat
P6	No	Yes	B	B	B	Not sure
P7	Yes	Reservation	A	A	A	Somewhat
P8	Reservations	Yes	B	A	B	No overlap
P9	Yes	Yes	B	Equal	B	Yes, clearly
P10	Reservations	Yes	A	Equal	A	No overlap

Table 5.23 Scenario 2 (RML Uplift) - Participant responses

P	Rep A Reuse	Rep B Reuse	Q2 Better	Q3 Easier	Q4 Confidence	Q5 Complementary
P1	Yes	Yes	B	A	B	Somewhat
P2	Yes	Yes	Equal	Equal	Equal	Somewhat
P3	Reservations	No	A	A	Neither	Not Sure
P4	Reservations	Yes	B	B	B	Somewhat
P5	Yes	Yes	B	A	B	Somewhat
P6	No	No	Neither	Equal	Neither	No overlap
P7	Yes	No	A	A	A	Somewhat
P8	Reservations	Yes	B	A	B	No overlap
P9	Yes	Yes	B	Equal	B	Yes clearly
P10	Reservations	Reservations	Equal	Equal	Equal	No overlap

Table 5.24 Scenario 3 (Ontology Alignment) - Participant responses

P	Rep A Reuse	Rep B Reuse	Q2 Better	Q3 Easier	Q4 Confidence	Q5 Complementary
P1	Yes	Yes	B	B	B	Somewhat
P2	Yes	Yes	B	Equal	Equal	Somewhat
P3	Yes	Yes	Equal	Equal	Equal	Somewhat
P4	Reservations	Yes	B	B	B	Somewhat
P5	Yes	Yes	B	A	A	No overlap
P6	Yes	Yes	Equal	A	A	No overlap
P7	Yes	No	A	A	A	Not sure
P8	Yes	Reservations	B	A	A	somewhat
P9	Yes	Yes	B	B	B	Yes clearly
P10	Yes	Yes	B	Equal	B	No overlap

Question 1: Would you consider this artefact reusable as a starting point?

Participants were asked to judge reusability based on provided metadata separately for each representation across three scenarios.

In **Scenario 1 (interlinking rule**), for Representation A (FAIR-IMPACT style), 6 out of 10 participants indicated they would consider reusing the artefact, 3 expressed reservations, and 1 declined. For Representation B (MMV), 7 participants indicated they would consider reusing the artefact, 2 expressed reservations, and 1 declined.

In **Scenario 2 (RML mapping)**, for Representation A (FAIR-IMPACT style), 5 out of 10 participants indicated they would consider reusing the artefact, 4 expressed reservations, and 1 declined. For Representation B (MMV), 6 participants indicated they would consider reusing the artefact, while 3 declined and 1 expressed reservations. The higher number of rejections under Representation B in this scenario reflects participants noting that several lifecycle fields were marked as "Not provided", which in some cases reduced rather than increased confidence.

In **Scenario 3 (ontology alignment),** for Representation A (FAIR-IMPACT style), 9 out of 10 participants indicated they would consider reusing the artefact, with 1 expressing reservations. For Representation B (MMV), 8 participants indicated they would consider reusing the artefact, with 1 expressing reservations and 1 declining. The strong performance of both representations in this scenario reflects the availability of concrete OAEI benchmark evaluation metrics, which supported reuse judgements regardless of representation.

Participants who expressed *yes but with reservations* were asked to briefly explain their reasoning. Six participants provided responses across the three scenarios.

For Representation A (FAIR-IMPACT style), reservations focused on insufficient detail. One participant noted that it was unclear how the mapping rules in Scenario 2 could be adapted for reuse. Another found that representation A lacked explanation of the rationale behind type restrictions and similarity thresholds in the interlinking scenario, noting that "B is much clearer on the WHY of the mapping logic, which A lacks." A third participant considered Representation A to have "much less detail and granularity" compared to B across all three scenarios.

For representation B (MMV), reservations were more varied. One participant raised concern that an identifier value in *mmv:hasIdentifier* "could not be found in the data", noting: "I become wary whenever I see values that seem to come from thin air." Another found it "difficult to find the connecting points" between ontology entities in the ontology alignment scenario. One participant found representation B valid but preferred the simpler, more widely known structure of A. In the uplift scenario, two participants noted that both representations answered the HOW but not the WHY of the mapping design.

- **Question 2: Which representation better supported your reuse decision?**

Across all three scenarios, Representation B (MMV) was most frequently selected as better supporting the reuse decision. For the interlinking scenario, 5 participants selected B, 4 selected A, and 1 rated both equally. For the uplift scenario, 5 selected

B, 2 selected A, 2 rated both equally, and 1 selected neither. In the ontology alignment scenario, 7 selected B, 1 selected A, and 2 rated both equally, the strongest result for Representation B across the study.

- **Question 3: Which representation was easier to understand?**

Understandability results were more evenly distributed. For interlinking scenario, 3 participants found B (MMV) easier, 3 found A (FAIR-IMPACT style) easier, and 4 rated both equally. For the uplift scenario, 5 participants found A easier, 1 found B easier, and 4 rated both equally. For the ontology alignment scenario, 4 participants found A easier, 3 found B easier, and 3 rated both equally. Representation A was rated as easier to understand in two out of three scenarios, suggesting a trade-off between the depth of information provided by B and ease of interpretation.

- **Question 4: Which representation gave you more confidence in your decision?**

For the interlinking scenario, 5 participants reported greater confidence with Representation B (MMV), 4 with Representation A, and 1 rated both equally. In the uplift scenario, 5 participants reported greater confidence with Representation B, 1 with Representation A, 2 rated both equally, and 2 selected neither both of whom noted that "Not provided" fields reduced the value of Representation B in this scenario. In the ontology alignment scenario, opinions were evenly split between Representation B (4), Representation A (4), and equal (2).

- **Question 5: Do you think the two metadata models complement each other?**

Across all three scenarios, 5 out of 10 participants selected "somewhat complementary", 3 selected "no, they mostly overlap", 1 selected "yes, clearly", and 1 selected "not sure". No participant selected "not compatible". This consistent pattern across all three scenarios suggests that participants perceived the two representations as serving distinct rather than redundant purposes.

- **Overall preference (Final question)**

After completing all three scenarios, 3 participants preferred Representation B (MMV) overall, 3 preferred using both representations together, 2 preferred Representation A (FAIR-IMPACT style), and 2 expressed no preference. Together, 6 out of 10 participants either preferred B or found both representations valuable, supporting the view that lifecycle-based metadata offers useful contextual information beyond what artefact-level metadata alone can provide.

- **Qualitative feedback**

Recurring themes in the open-ended responses included the value of the lifecycle structure of Representation B (MMV) for explaining the purpose, assumptions, and design rationale of mappings information participants noted was absent from Representation A. However, four participants independently raised concerns about the use of string-valued properties in Representation B (MMV), noting that their usefulness depends entirely on the quality and completeness of the information provided. Two participants noted the practical burden of completing all lifecycle fields, with some participants suggesting that "Not provided" values reduced confidence rather than conveying neutral absence. A recurring suggestion across participants was that the two representations work best together, Representation A for quick initial assessment and Representation B for in-depth reuse evaluation. Table 5.25 shows themes from qualitative feedback (n = 10)

Table 5.25 Qualitative feedback themes

Theme	Quotes	n
Lifecycle structure aids navigation	"The 'think aloud' in going from Analysis to Design to Development"	4
String values reduce trust	"I become wary whenever I see values that seem to come from thin air"	4
Rep A better for quick assessment	"A is great for a quick assessment. B is great for more detail"	3
Representations complement each other	"They complement each other even though there are some overlaps"	3
Practical burden of completing fields	"My concern is getting people to actually fill it"	2
Missing WHY of design decisions	"B is much clearer on the WHY of the mapping logic, which A lacks"	2

5.6.8 Discussion

The results provide evidence across all three hypotheses, with findings that both support and qualify the central claims of this study.

Hypothesis 1, that participants would be better able to judge reuse when using lifecycle-based metadata, is supported. Representation B (MMV) was consistently selected as the representation that better supported the reuse decision across all three scenarios, most strongly in the ontology alignment scenario where 7 out of 10 participants preferred B. Reuse judgements under Representation B were also slightly more positive in the interlinking and uplift scenarios. However, the benefit was context-dependent. In the uplift scenario, the presence of several "Not provided" fields in Representation B resulted in more rejections compared to Representation A (FAIR-IMPACT style), indicating that incomplete lifecycle metadata can reduce rather than improve perceived reusability. This suggests that the value of lifecycle-based metadata is directly tied to the completeness and quality of the information provided, which in turn reinforces the importance of tool support to guide users in completing metadata fields consistently.

Hypothesis 2, that participants would prefer Representation B (MMV), received mixed but broadly positive support. While B was preferred by the largest single group (3 participants), an equal number preferred using both representations together, and 4 participants chose A (FAIR-IMPACT style) or expressed no preference. The most consistent finding across all three scenarios was complementarity, 6 out of 10 participants rated the two representations as somewhat or clearly complementary. This suggests that lifecycle-based metadata is best positioned as an extension of rather than a replacement for concise artefact-level metadata. Representation A was valued for quick initial assessment and filtering, while Representation B provided the contextual depth needed for more informed reuse decisions. This finding is consistent with the design intent of

LMMD, which was never intended to replace existing approaches but to complement them with lifecycle context.

Hypothesis 3, that Representation B (MMV) would provide greater decision confidence, is supported in two out of three scenarios. Representation B led on confidence in the interlinking and uplift scenarios, with 5 participants each selecting B. In the ontology alignment, however, results were evenly split between B and A (FAIR-IMPACT style). This can be explained by the nature of the scenario itself the OAEI anatomy alignment provided concrete benchmark evaluation metrics including precision, recall, F-measure, and coherence, which gave participants direct quantitative evidence of mapping quality under both representations. In scenarios where such objective evidence was not available, the lifecycle context provided by B played a more significant role in building confidence.

A noteworthy finding concerns understandability. Despite Representation B (MMV) being significantly longer and containing more metadata fields than Representation A, it received equal or better understandability ratings in several scenarios most notably in the interlinking scenario, where understandability was split equally between A, B, and equal. This suggests that the lifecycle phase structure of MMV, which organises metadata into clearly labelled groups Analysis, Design, Development, Testing, and Maintenance supports quick orientation even when the total volume of information is higher. Rather than presenting metadata as a flat list of fields, the phase-based grouping allows users to navigate directly to the information most relevant to their reuse decision. This is an important design insight: structured organisation of metadata can partially compensate for greater information density, making lifecycle-based documentation more accessible than its length alone would suggest.

A recurring concern across qualitative responses was the use of string-valued properties in Representation B (MMV). Several participants noted that free-text values are difficult to verify, and that their usefulness depends entirely on the care

of the person who populated them. One participant explicitly stated that seeing an identifier value that "could not be found in the data" made them wary of relying on the metadata. This finding directly reinforces the refinement identified in Study 3 that transitioning from string-valued to object properties linked to controlled vocabularies would improve the semantic precision and verifiability of lifecycle metadata. The practical burden of completing all lifecycle fields was also raised by multiple participants, underlining the importance of guided annotation tools and template support to reduce documentation overhead.

The qualitative feedback also highlighted an important gap: several participants noted that both representations answered the HOW of a mapping but not sufficiently the WHY the rationale behind design decisions. This is precisely what properties such as *mmv:hasJustification* and *mmv:hasDesignDecision* are intended to capture, suggesting that clearer labelling and richer examples for these fields would improve their perceived value in practice.

5.6.9 Experiment Summary

Study 4 investigated whether lifecycle-based metadata supports reuse decision-making across three mapping types. Ten participants compared Representation A (FAIR-oriented artefact-level metadata) and Representation B (MMV) across three scenarios.

All three hypotheses were supported. Representation B was consistently selected as better supporting reuse decisions, most strongly in the ontology alignment scenario (7/10). Decision confidence favoured B in two of three scenarios. The most consistent finding was complementarity, 6 out of 10 participants rated the two representations as complementary, with 6 out of 10 preferring B alone or both together.

Qualitative feedback identified two concerns: the limitation of string-valued properties, and the practical burden of completing all lifecycle fields, which underlines the importance of guided annotation support.

The findings position lifecycle-based metadata as a complement to rather than a replacement for artefact-level metadata, consistent with the design intent of LMMD.

5.7 Chapter Summary

This chapter presented four evaluation studies designed to assess different aspects of the Lifecycle-Based Metadata Model for Declarative Mappings (LMMD) and its associated artefacts.

Study 1, a community survey, gathered feedback from mapping practitioners across experience levels to assess the perceived relevance and completeness of the proposed lifecycle metadata fields. The findings validated the relevance of the model and identified analysis-phase fields as highest priority, informing the design of LMMD and the subsequent evaluation studies.

Study 2, a usability evaluation of MetaSEMAP, assessed the tool's effectiveness and user satisfaction across three mapping types. The results confirmed sufficient usability across all conditions, with structured metadata annotation feasible for uplift mappings, ontology alignments, and interlinking specifications. Participants showed a clear preference for the Named Graph metadata representation over RDF-star.

Study 3, an ontology evaluation, examined the conceptual clarity, structural coherence, and vocabulary reuse of MMV through expert review and automated validation. The evaluation confirmed that MMV conforms to recognised ontology engineering principles. Expert feedback informed several refinements incorporated into MMV v0.2, including the addition of hasValidationReport and updated documentation.

Study 4, a mapping reuse experiment, investigated whether lifecycle-based metadata supports more informed reuse decisions. The results showed that

lifecycle-based metadata consistently better supported reuse decisions than the FAIR-IMPACT metadata, and that the two representations were perceived as complementary rather than redundant. The phase-based structure of MMV was found to improve navigability even at higher information density.

The use case study on metadata-guided LLM mapping generation is presented separately in Chapter 6, where it is framed as a demonstration of a potential application of LMMD beyond its original evaluation scope.

Overall, the four studies provide quantitative and qualitative evidence that LMMD supports mapping documentation, usability, formal correctness, and reuse decision-making. The use case presented in Chapter 6 provides additional evidence that LMMD supports AI-assisted generation. The cross-study coherence of findings strengthens the claim that a lifecycle-based metadata model addresses a genuine and persistent gap in the declarative mapping community.

Chapter 6: Use Case - Metadata-Guided LLM Mapping Generation

This section presents a controlled experiment evaluating whether lifecycle-based metadata improves the quality of LLM-generated declarative mappings. The study compares outputs generated with and without MMV lifecycle metadata across six real-world scenarios covering three mapping types: RML uplift mappings, ontology alignment, and interlinking. The section is structured as follows: Motivation; Hypotheses; Methodology; Experiment Structure; Results; Experiment Summary; and Discussion.

6.1 Motivation

With the increasing use of Large Language Models (LLMs) for generating declarative mappings, it was considered important to assess whether structured lifecycle-based metadata can improve automated outputs.

A review of LLM-based mapping generation approaches in Section 3.3 revealed a recurring pattern: metadata-like information was frequently embedded within natural language prompts in an ad hoc way. For example, stakeholder roles were used to frame alignment tasks, input format descriptions were included to constrain output structure, and domain assumptions were stated informally to guide vocabulary selection. However, these fields appeared without a common vocabulary or lifecycle framework. Crucially, the fields used in those prompts correspond closely to fields already defined in LMMD, which was originally designed for mapping documentation and reuse rather than generation. This observation motivated the central question of Study 5: whether LMMD, structured according to lifecycle phases, could serve as a reusable template for guiding LLM mapping generation in a more correct and complete way.

An early evaluation of this idea was presented at the KG-NeSy 2025 workshop co-located with SEMANTiCS 2025 [72], where the approach was discussed with the Semantic Web community and received valuable feedback. That evaluation used

synthetic scenarios and a limited prompt structure. The study described here extended that evaluation by applying the approach to six real-world mapping tasks across all three mapping types supported by LMMD. The metadata fields used to guide the LLM prompts were selected from LMMD based on two criteria: relevance to the generation task as identified in Study 1, and alignment with the types of contextual information shown to influence generation quality in the literature reviewed in Section 3.3. This study therefore investigated whether structured lifecycle-based metadata might improve the semantic correctness and completeness of LLM-generated mappings across different mapping types.

6.2 Hypothesis

The following hypotheses were defined for this experiment:

- **Hypothesis 1 (Generation quality):** LLM-generated mappings produced with lifecycle-based metadata will be more semantically correct than mappings generated without such metadata.
- **Hypothesis 2 (Generation completeness):** LLM-generated mappings produced with lifecycle-based metadata will be more complete than mappings generated without such metadata.
- **Hypothesis 3 (Cross-type consistency):** A consistent improvement in generation quality due to lifecycle-based metadata will be observed across different declarative mapping types, including uplift mappings, ontology alignment, and interlinking.

These hypotheses extend the findings of Studies 1 through 4 by evaluating whether lifecycle-based metadata can function as structured prompt guidance to improve AI-assisted mapping generation.

6.3 Methodology

To test these hypotheses, a controlled experiment was conducted to compare LLM-generated mappings under two conditions: (i) without lifecycle-based metadata and (ii) with lifecycle-based metadata. For each task, the same input data and task description were provided to the LLM, with the only difference being the inclusion of lifecycle-based metadata in the prompt.

The lifecycle-based metadata included in the prompt condition was derived from the findings of Study 1. In that study, practitioners evaluated the relevance of LMMD metadata fields for mapping discovery and mapping quality assessment. For the purposes of Study 5, fields that received strong relevance ratings from expert and intermediate participants were shortlisted, and a subset of these was selected for prompt construction based on their expected usefulness for guiding mapping generation. The selected fields focused on generation-relevant aspects such as mapping requirement, mapping type, domain, assumptions, technical requirements, source characteristics, design decisions, justifications, mapping method, and testing expectations. Metadata fields primarily related to publication and some development details were not included, as they were not expected to directly influence generation quality.

The experiment covered three types of declarative mappings: uplift mapping (RML), ontology alignment, and interlinking. For each mapping type, scenarios were selected based on real-world or realistic use cases used throughout this thesis. For each scenario, the LLM was executed twice: once without metadata (baseline condition), and once with the selected lifecycle-based metadata (enhanced condition). In the enhanced condition, the metadata was provided as structured contextual input organised according to LMMD phases, including analysis, design, development, testing, and where relevant maintenance. Listings 5.1 and 5.2 show illustrative example for both the baseline and guided prompt structures respectively.

Listing 6.1 MMV Baseline prompt structure

```
1. System: You are a helpful assistant that generates valid RML
2. mappings from structured task descriptions. Return only valid
3. RML mapping in Turtle syntax. Do not include explanations,
4. comments, markdown, or extra text.
5.
6. User input: Generate a RML mapping artefact for the following task.
7.
8. Task description:
9. You are given a CSV dataset with the following columns:
10. RegionName, RegionCode, ParentRegion, Latitude, Longitude.
11.
12. Generate an RML mapping to:
13. - represent each region as a resource
14. - use rdfs:label for region name
15. - use skos:Concept as class
16. - represent hierarchy using skos:broader
17. - include geo:lat and geo:long
18.
19. Return valid RML in Turtle syntax.
```

Listing 6.2 Guided prompt structure (with lifecycle metadata)

```
1. System: You are a helpful assistant that generates valid RML
2. mappings from structured task descriptions. Return only valid
3. RML mapping in Turtle syntax. Do not include explanations,
4. comments, markdown, or extra text.
5.
6. User input: Generate a RML mapping artefact for the following task.
7.
8. Task description:
9. You are given a CSV dataset with the following columns:
10. RegionName, RegionCode, ParentRegion, Latitude, Longitude.
11.
12. Generate an RML mapping to:
13. - represent each region as a resource
14. - use rdfs:label for region name
15. - use skos:Concept as class
16. - represent hierarchy using skos:broader
17. - include geo:lat and geo:long
18.
19. Lifecycle metadata:
20. Requirement: Transform regional CSV data into RDF with
21. hierarchy and coordinates
22. Mapping type: RML
23. Domain: Geography
24. Assumptions: RegionCode uniquely identifies each region;
25. ParentRegion refers to another region
26. Technical requirements: Output must be valid RML in Turtle
27. syntax
```

```
Source: CSV dataset with region information
Input format: CSV

Design decisions:
- represent regions as skos:Concept
- use rdfs:label for names
- use skos:broader for hierarchy
- map coordinates using geo:lat and geo:long

Justifications:
- SKOS supports hierarchical concepts
- geo properties represent spatial coordinates

Quality metrics:
- all fields must be mapped
- correct hierarchy representation
- valid URI generation

Mapping method: schema-driven mapping
Mapping format: RML

Testing expectations:
- each row produces one resource
- no missing attributes
- mapping is syntactically valid

Return valid RML in Turtle syntax.
```

Full prompts and generated outputs for all six scenarios are available at https://github.com/SarahAlzahranitcd/Thesis-Materials/tree/main/Study-5-LLM-Generation

A simple web-based interface was developed to run the experiment (Figure 5.3). For each run, an identifier was assigned, the mapping type was selected, and a run number was used to distinguish between executions, run 1 for the baseline and run 2 for the guided condition. The task description was entered first, followed by the metadata when the enhanced condition was selected. The full prompts and generated outputs were automatically logged and stored in a CSV file for analysis. The experiments were conducted using the GPT-4o-mini model at temperature 0 to ensure deterministic and reproducible outputs. The interface is accessible through the **MetaSEMAP**[10] main interface. The goal of this experiment was not

[10] https://metasemap-1-sarah-alzahrani.replit.app/llmexperiment

to benchmark LLM performance, but to evaluate the effect of structured metadata on generation quality under controlled conditions.

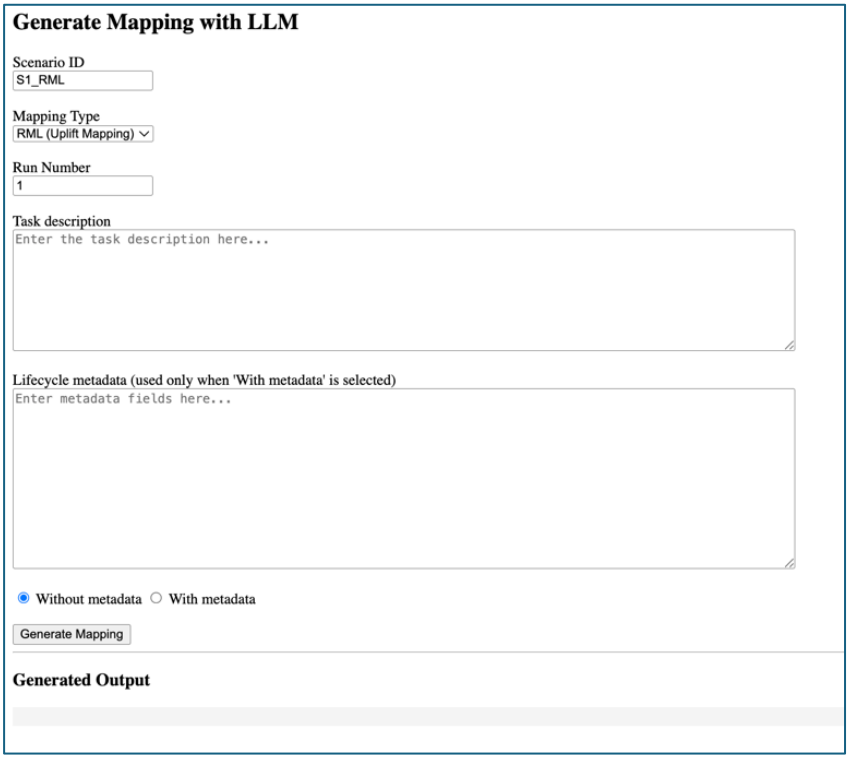


Figure 6.1 Web-based interface used to run the LLM mapping generation experiment

6.4 Experiment Structure

The experiment was structured as follows:

1. **Scenario preparation:** Six scenarios were defined across three mapping types: two RML uplift mappings, two ontology alignments, and two interlinking specifications. The six scenarios were: S1A (ERA railway RML mapping)[8], S1B (GTFS transit stops RML mapping)[11], S2A (OAEI Anatomy alignment)[12], S2B (OAEI Conference alignment)[13], S3A (VRTI Irish places interlinking)[14],

[11] https://data.gov.ie
[12] https://oaei.ontologymatching.org/2024/results/anatomy/index.html
[13] https://oaei.ontologymatching.org/2024/results/conference/index.html
[14] https://virtualtreasury.ie

and S3B (DBpedia-GeoNames interlinking) [31]. All scenarios were grounded in real-world mapping tasks.

2. **Prompt construction:** For each scenario, two prompts were prepared: a baseline prompt containing only the task description and input data and an enhanced prompt including a structured subset of lifecycle-based metadata model (LMMD). The metadata fields used to guide the LLM prompts were selected from LMMD based on two criteria: relevance to the generation task as identified in Study 1, and alignment with the types of contextual information shown to influence generation quality in the literature reviewed in Section 3.3.

3. **LLM execution:** The LLM was executed for each scenario under both conditions, generating mapping outputs. Steps 1 and 2 were repeated for each scenario before execution.

4. **Output evaluation:** Generated outputs were compared and scored based on correctness and completeness. Each generated output consisted of two components. The first was the declarative mapping specification itself: an RML mapping in Turtle syntax for uplift scenarios, an Alignment API RDF/XML document for alignment scenarios, and a SPARQL CONSTRUCT query for interlinking scenarios. The second component, produced only in the enhanced condition, was an MMV metadata annotation block appended directly to the mapping output. This block was generated by the LLM itself based on the metadata provided in the prompt, and represented a structured annotation of the produced mapping artefact using MMV properties. This output: mapping specification plus lifecycle metadata annotation demonstrates that LMMD can guide not only the generation of the mapping itself but also the simultaneous production of its metadata documentation.

5. **Result comparison:** Scores were analysed to assess the effect of lifecycle metadata across different mapping types. Each output was evaluated against a

set of scenario-specific criteria covering two dimensions: correctness, assessing whether the generated output accurately represented the intended semantics including appropriate vocabulary selection, valid structure, and absence of fabricated values; and completeness, assessing whether all required mapping elements were present. Each criterion was scored independently using a binary scheme (1 = present and correct, 0 = missing or incorrect). In one instance (S1B stop_url), a score of 0.5 was assigned to reflect a partially correct implementation where the intent was correct but the syntax was invalid. All evaluation was conducted manually by the author of this thesis against the task requirements defined for each scenario. This scoring approach follows a task-specific, criteria-based evaluation strategy consistent with recent work on LLM-generated mappings. Hofer et al. [19] argue that standard metrics are insufficient for evaluating mapping generation quality and advocate for custom evaluation measures tailored to the specific mapping task. The evaluation in this study adopts the same principle, using scenario-specific correctness and completeness criteria to assess the fitness of generated mappings relative to their task requirements rather than relying on generic benchmarks.

To complement the manual evaluation, syntactic validation was performed automatically on each generated output using the web-based interface developed for the experiment. For RML outputs, the generated Turtle was parsed using the rdflib Python library and the number of valid triples was recorded. For ontology alignment outputs, the generated RDF/XML was parsed using the same library. For interlinking outputs, the generated SPARQL query was checked for the presence of required structural keywords including CONSTRUCT, WHERE, and *owl:sameAs*. Outputs that failed to parse were flagged as syntactically invalid. This automated validation confirmed that guided condition outputs were syntactically valid in four of the six scenarios. In S1B a minor RML syntax error was identified in the stop_url mapping, and in S2A and S2B the alignment outputs had structural issues including missing declarations, which are noted in the scenario results in Section 5.7.5.

6.5 Results

The results are presented across six scenarios grouped by mapping type: two RML uplift mappings (S1A and S1B), two ontology alignments (S2A and S2B), and two interlinking specifications (S3A and S3B). For each scenario, the baseline and guided prompts are presented alongside the generated outputs and evaluation scores.

A. RML Uplift Mappings

Scenario S1A: ERA Railway RML Mapping

The ERA scenario evaluated the effect of lifecycle-based metadata on LLM-generated RML mappings for European railway infrastructure data. The task involved generating an RML mapping from a CSV dataset with six columns: *OperationalPointName, OperationalPointType, Latitude, Longitude, CountryCode,* and *LineIdentification*. The evaluation criteria and results are presented in Tables 5.26 and 5.27.

Table 6.1 Evaluation criteria for S1A and S1B

Criterion	Description
Syntactic validity	Valid RML structure that parses correctly
All 6 columns mapped	All columns present in the output
Established vocabulary used	Uses recognised transport or geographic vocabularies
URI construction from OperationalPointName	Persistent URI template correctly applied
CountryCode as typed IRI resource	CountryCode represented as IRI not literal
LineIdentification as typed IRI resource	LineIdentification represented as IRI not literal
Coordinate datatype xsd:decimal	Coordinates assigned correct datatype
MMV metadata block appended	MappingArtifact annotation block present

Table 6.2 S1A results (Correctness / Completeness)

Condition	Correctness	Completeness	Key observation
Baseline (no metadata)	3/8	6/6	Generic ex: vocabulary throughout; nested predicateObjectMap structure invalid; CountryCode and LineIdentification as literals
Metadata Guided	8/8	6/6	schema.org and geo:wgs84 vocabularies adopted; CountryCode and LineIdentification as typed IRI resources; MMV metadata block appended

Both conditions produced complete mappings covering all six columns. The baseline (without metadata in the prompt context) used generic *ex: vocabulary* throughout and introduced an invalid nested *predicateObjectMap* structure. *CountryCode* and *LineIdentification* were mapped as plain literals rather than typed resources. The MMV metadata guided condition adopted schema.org and geo:wgs84_pos vocabularies, represented *CountryCode* and *LineIdentification* as typed IRI resources using URI templates, and appended a complete MMV metadata block. The improvements were directly traceable to specific metadata fields: *mmv:hasRiskOrIssue* warning against generic vocabulary caused the adoption of established vocabularies, and *mmv:hasDomainAssumption* specifying that *CountryCode* and *LineIdentification* should be typed resources caused the change from literals to IRIs.

Scenario S1B : GTFS Transit Stops RML Mapping

The GTFS scenario evaluated the effect of lifecycle-based metadata on RML mapping generation for public transit stop data. The task involved generating an RML mapping from a GTFS stops.txt file with six columns: *stop_id, stop_name, stop_lat, stop_lon, stop_url,* and *location_type*. The evaluation criteria and results are presented in Tables 5.26 and 5.28.

Table 6.3S1B results (Correctness / Completeness)

Condition	Correctness	Completeness	Key observation
Baseline (no metadata)	4/8	6/6	Generic ex: vocabulary; stop_url as plain literal; location_type as raw integer literal
Metadata Guided	7.5/8	6/6	transit: and geo:wgs84 vocabularies adopted; location_type as typed classification; stop_url attempted as IRI with minor syntax error; MMV block appended

The MMV metadata guided condition adopted the transit: vocabulary from vocab.org and geo:wgs84 for coordinates, directly responding to *mmv:hasRiskOrIssue* and *mmv:hasDesignDecision*. The *location_type* column was represented using *transit:LocationType* as a typed classification rather than a raw integer literal, directly responding to *mmv:hasDomainAssumption*. The *stop_url* column was attempted as a typed IRI with a minor syntax error, the intent was correct but the RML property used was incorrect (*rr:datatype rml:IRI* instead of *rr:termType rr:IRI*). The MMV metadata block was appended correctly. The improvement pattern seen is consistent with S1A lifecycle metadata from the Analysis and Design phases directly influenced vocabulary selection and semantic modelling decisions.

B. Ontology Alignment

Scenario S2A: OAEI Anatomy Alignment

The OAEI Anatomy scenario evaluated the effect of lifecycle-based metadata on LLM-generated ontology alignments between mouse anatomy (MA) and human anatomy (NCI Thesaurus) ontologies. Neither condition was provided with the actual ontology files. This design decision was deliberate, providing ontology files

only to the guided condition would confound the metadata variable with a data provision variable, making it impossible to attribute observed improvements specifically to LMMD. Table 5.29 shows the evaluation criteria for S2A and S2B

Table 6.4 Evaluation criteria for S2A and S2B

Criteria	Description
Valid Alignment API format	Correct RDF/XML structure with required declarations
Real ontology URIs used	Uses actual MA_ and NCI_ namespace patterns
Correct relation symbol	Uses = not the word equivalent
Varied confidence values	Confidence values reflect degree of certainty not uniform 1.0
align:Cell structure present	Correspondence cells correctly structured
MMV metadata block appended	MappingArtifact annotation block present

Table 6.5 S2A results (Correctness / Completeness)

Condition	Correctness	Completeness	Key observation
Baseline (no metadata)	1/6	1/6	Fabricated URIs using http://example.org/ontology/mouse#, scores zero against OAEI gold standard; relation symbol incorrect; all confidence values 1.0
Metadata Guided	4.5/6	4/6	Real MA_ and NCI_ namespaces adopted; correct relation symbol =; varied confidence values; align:Cell wrapper missing

The S2A scenario produced the most dramatic improvement across all scenarios. The baseline condition generated completely fabricated URIs that cannot be evaluated against the OAEI gold standard and are unusable in practice. The MMV metadata guided condition adopted real MA_ and NCI_ URI schemes directly from *mmv:hasDomainAssumption,* changed the relation symbol from the invalid word equivalent to the correct = directly from *mmv:hasTechnicalRequirement*, and produced varied confidence values directly from *mmv:hasDesignDecision*. Two

limitations remained: the *align:Cell* wrapper was missing, and specific numeric identifiers beyond the first example were sequential approximations rather than drawn from actual ontology content. This finding supports the argument that MMV lifecycle metadata can prevent hallucinated URIs and enforce structural quality, but cannot substitute for direct ontology content access when precise alignment is required.

Scenario S2B: OAEI Conference Alignment

The OAEI Conference scenario evaluated alignment generation between the *cmt* ontology and the *conference* ontology, both describing academic conference organisation. As with S2A, neither condition was provided with the actual ontology files. The evaluation criteria and results are presented in Tables 5.29 and 5.31.

Table 6.6 S2B results (Correctness / Completeness)

Condition	Correctness	Completeness	Key observation
Baseline (no metadata)	1.5/6	3/6	Fabricated http://example.org/cmt# URIs; relation equivalent used; all confidence values 1.0
Metadata Guided	5.5/6	5/6	Real http://cmt# and http://conference# namespaces adopted; correct relation symbol; varied confidence; naming distinction preserved; MMV block appended

The MMV metadata guided condition adopted the real *http://cmt*# and *http://conference*# namespaces directly from *mmv:hasDomainAssumption*, producing correspondences evaluatable against the OAEI gold standard. The guided output correctly distinguished between *cmt:ProgramCommittee* and *conference:Programme_committee,* preserving the naming difference between the two ontologies and demonstrating semantic awareness beyond simple label

matching. The serialisation format switched from RDF/XML to Turtle despite the task requesting RDF/XML, which represents a minor issue. Combined with S2A, both alignment scenarios demonstrate that *mmv:hasDomainAssumption* is the single most impactful metadata field for alignment generation, it prevents URI hallucination, which is the most damaging failure mode in LLM-generated alignments.

C. Interlinking

Scenario S3A: VRTI Irish Places Interlinking

The VRTI scenario evaluated interlinking query generation between Ordnance Survey Ireland (OSi) place records and Wikidata Irish places. The task involved generating a SPARQL CONSTRUCT query linking matching places using *owl:sameAs*. Neither condition was provided with RDF dataset files. Table 5.32 shows the evaluation criteria for S3A and S3B

Table 6.7 Evaluation criteria for S3A and S3B

Criteria	Description
Valid SPARQL structure	CONSTRUCT query syntactically valid
owl:sameAs present	Correctly used in CONSTRUCT
Case-insensitive name matching	Uses lcase() for name comparison
Coordinate tolerance with ABS()	Uses ABS() with tolerance rather than exact equality
Minimum two matching signals	At least two independent signals used
Tolerance value explicitly stated	Specific tolerance value present in query
MMV metadata block appended	MappingArtifact annotation block present

Table 6.8 S3A results (Correctness / Completeness)

Condition	Correctness	Completeness	Key observation
Baseline (no metadata)	3/7	4/6	Exact string matching; exact coordinate equality; no tolerance defined
Metadata Guided	7/7	6/6	lcase() name matching; ABS() <= 0.05 coordinate tolerance; MMV metadata block appended

Two improvements were directly traceable to specific MMV fields. Case-insensitive matching using lcase() was adopted directly from *mmv:hasDomainAssumption* specifying that case-insensitive comparison is required. Coordinate tolerance ABS(?latA - ?latB) <= 0.05 was adopted directly from *mmv:hasDomainAssumption* specifying a tolerance of 0.05 decimal degrees. The exact tolerance value was adopted precisely from the metadata. Both baseline failures: exact string matching and exact coordinate equality would fail in practice for real Irish place data due to spelling variants and coordinate precision differences between OSi and Wikidata. A known error in both conditions is the use of ?placeB a *wdt:P17*, where *wdt:P17* is a Wikidata property for country rather than a class. This error reflects a limitation in the task prompt description and appears equally in both conditions, so it does not affect the comparative evaluation.

Scenario S3B: DBpedia-GeoNames Interlinking

The DBpedia-GeoNames scenario evaluated interlinking query generation between DBpedia geographic places and GeoNames features. This scenario introduced a critical additional challenge: DBpedia represents country as a URI resource such as dbr:Ireland, while GeoNames represents country as an ISO 3166-1 alpha-2 string such as IE. These are incompatible types that will silently produce zero results if compared directly. Table 5.34 shows S3B results for both correctness and Completeness.

Table 6.9 S3B results (Correctness / Completeness)

Condition	Correctness	Completeness	Key observation
Baseline (no metadata)	3/7	5/6	Exact string matching; exact coordinate equality; broken country comparison that silently returns zero results
Metadata Guided	7/7	6/6	lcase() matching; ABS() tolerance 0.05; country signal excluded entirely; MMV metadata block appended

Two further improvements were directly traceable to specific MMV fields. The guided condition omitted country as a matching criterion entirely, directly responding to mmv:hasAssumption which warned that DBpedia represents country as a URI resource (e.g. dbr:Ireland) while GeoNames represents it as an ISO string (e.g. IE), making direct comparison impossible. This is the most instructive finding across all interlinking scenarios: lifecycle metadata captures domain-specific knowledge that cannot be inferred from schema information alone. The incompatibility of country representations across datasets is knowledge that belongs to the Analysis phase of the mapping lifecycle and is precisely what *mmv:hasAssumption* is designed to record. The baseline country comparison failure would not be caught by syntax validation, it is a semantic error that would only become apparent when the query returns zero results in practice. The guided output also adopted *lcase()* for name matching and *ABS()* with tolerance 0.05 for coordinate matching, both directly traceable to *mmv:hasDomainAssumption*. The MMV metadata block was appended correctly.

6.5.1 Discussion

The results provide consistent evidence that lifecycle-based metadata can improve LLM-generated mapping quality across all three mapping types. This section discusses the findings in relation to the three hypotheses, the key design choices made in the experiment, and the broader implications for the role of LMMD in AI-assisted mapping generation.

Hypothesis 1 (Generation quality). The results provide evidence that Hypothesis 1 is supported. Semantic correctness improved in all six scenarios. The most substantial improvements were observed in the alignment scenarios, where the baseline (without metadata in prompt context) produced completely fabricated ontology URIs that scored zero against the OAEI gold standard. The MMV metadata guided condition adopted real ontology namespaces, correct relation symbols, and varied confidence values which are all directly traceable to specific MMV metadata fields. In the interlinking scenarios, the guided condition prevented a silent correctness failure in scenario S3B where the baseline attempted to compare incompatible country representations across DBpedia and GeoNames. This type of failure cannot be detected by syntax validation and would only become apparent at query execution time.

Hypothesis 2 (Generation completeness). The results provide evidence that Hypothesis 2 is partially supported. In four of the six scenarios the baseline (without metadata in prompt context) already achieved high completeness, confirming that task descriptions alone are sufficient for structural coverage. The benefit of lifecycle metadata was therefore not primarily in adding missing elements, but in improving the semantic quality and correctness of the elements already present. This distinction is important: the contribution of LMMD is not to tell the LLM what to include, but to tell it how to generate declarative mappings correctly.

Hypothesis 3 (Cross-type consistency). The results provide evidence that Hypothesis 3 is supported. Improvements were observed consistently across RML uplift mappings, ontology alignments, and interlinking specifications, confirming that the benefit of lifecycle-based metadata generalises across mapping types. The most impactful fields were found to be *mmv:hasDomainAssumption, mmv:hasAssumption, mmv:hasRiskOrIssue,* and *mmv:hasDesignDecision*.

Design choice: why mapping specifications and not full mapping files. A key design choice in this study was to evaluate the quality of the generated mapping

specifications themselves: the RML Turtle file, the Alignment API document, the SPARQL query rather than executing them against real data and measuring the quality of the resulting RDF output. This choice was deliberate. The goal of the study was to assess whether MMV metadata improves the LLM's modelling decisions: vocabulary selection, URI construction, structural validity, and semantic correctness. These properties are directly observable in the mapping specification and do not require execution. The mapping specification is also the primary artefact that practitioners share, reuse, and maintain. Evaluating its quality directly is therefore more relevant to the thesis claim than evaluating downstream execution results.

Design choice: why metadata-only prompts without additional files. A second key design choice was to provide the LLM with metadata only, no ontology files, no source data samples, no schema definitions in both conditions. This design decision was deliberate with the intention of isolating the specific contribution of LMMD to generation quality. The experiment models a realistic practitioner scenario: one where lifecycle metadata is available because it was produced during the mapping project, but where the practitioner does not necessarily have access to the full source data or ontology at generation time. This is precisely the scenario LMMD is designed for supporting reuse and generation from documented metadata rather than requiring access to the original artefacts. Future work could explore the combined effect of metadata and source files, but this is outside the scope of the present evaluation.

Implications for LMMD as a generation template. The findings (based on MMV instantiation of LMMD) suggest that LMMD serves a dual purpose. It was originally designed for mapping documentation and reuse, but the results of this study demonstrate that it also functions as a structured generation template. The fields that had the greatest impact: domain assumptions, technical constraints, design decisions, and risk statements are precisely the fields that encode practitioner knowledge that is not inferable from the task description alone. This positions LMMD not only as a documentation model but as a bridge between

human expertise and automated generation, encoding the contextual knowledge that guides good mapping practice in a form that an LLM can act on directly.

Limitations. The evaluation was conducted using a single LLM (GPT-4o-mini) at temperature 0. Results may differ across models or prompt formulations. Syntactic validation of generated outputs was performed automatically using rdflib parsing for RML and alignment outputs and keyword-based checks for SPARQL queries. Semantic evaluation including vocabulary appropriateness, URI quality, and structural correctness was conducted manually by the author, which introduces subjectivity in borderline scoring decisions. These limitations do not undermine the core finding that lifecycle metadata can consistently improve semantic correctness, but they indicate that the approach requires further evaluation at larger scale and with additional models.

6.5.2 Experiment Summary

This study evaluated whether MMV lifecycle metadata improves the quality of LLM-generated declarative mappings across six real-world scenarios covering three mapping types. In all six scenarios, the guided condition outperformed the baseline. The most impactful metadata fields were *mmv:hasDomainAssumption, mmv:hasAssumption, mmv:hasRiskOrIssue,* and *mmv:hasDesignDecision*, all from the Analysis and Design phases of LMMD. These fields encode practitioner knowledge that cannot be inferred from the task description alone, including domain-specific URI patterns, coordinate precision constraints, and incompatible data representations across datasets.

The findings provide evidence to support all three hypotheses and demonstrate that lifecycle-based metadata can function as a guided template for LLM-assisted mapping generation, improving semantic correctness and structural quality across different mapping types. Study 5 also demonstrates a secondary contribution: the guided condition produced an MMV metadata block alongside each mapping, showing that LMMD can guide not only the generation of the mapping itself but also the simultaneous production of its metadata documentation.

Chapter 7: Conclusion and Future Work

This chapter draws conclusions from the research presented in this thesis. Section 7.1 discusses the extent to which the research objectives were fulfilled. Section 7.2 presents the research contributions. Section 7.3 describes the impact and uptake of the research. Section 7.4 presents future work directions. Section 7.5 provides final remarks

7.1 Research objectives

The extent to which the research objectives were achieved is discussed in this section. The research question of this thesis was:

To what extent can a lifecycle-based metadata model improve the documentation, reuse, and generation of declarative mappings?

Research Objective 1, RO1 was to identify and validate comprehensive metadata requirements for declarative mappings across different mapping types and lifecycle phases. This objective was achieved through a review of existing mapping lifecycle models and metadata standards presented in Chapters 2 and 3, followed by a community survey (Study 1). The survey confirmed that lifecycle metadata fields are perceived as highly relevant across experience levels, with analysis-phase fields rated most consistently. The results were used to establish the requirements for LMMD and informed the design of all subsequent studies.

Research Objective 2, RO2 was to design a lifecycle-based metadata model and implement it as an ontology. This objective was achieved through the design of LMMD and its formalisation as MMV, described in Chapter 4. MMV was evaluated using competency questions, expert review (Study 3), and automated validation using OOPS! and FOOPS!, all of which confirmed conceptual clarity, lifecycle coverage, and conformance to ontology engineering best practices.

Research Objective 3, RO3 was to develop MetaSEMAP as a tool demonstrating the operationalisation of LMMD. MetaSEMAP was evaluated in Study 2 with 50 participants across three mapping types. The results confirmed that the tool is usable by non-expert users, with an overall PSSUQ score of 2.49 below the 3.5 sufficiency threshold, and that lifecycle metadata annotation is feasible across all three supported mapping types.

Research Objective 4, RO4 was to evaluate LMMD through usability testing and a mapping reuse experiment. The usability evaluation (Study 2) confirmed sufficient usability across all conditions. The mapping reuse experiment (Study 4) compared a FAIR-oriented representation and the MMV lifecycle-based representation across ten expert participants and three mapping scenarios. Representation B was consistently selected as better supporting reuse decisions, with 6 out of 10 participants rating the two representations as complementary.

Research Objective 5, RO5 was to examine the potential impact of lifecycle-based metadata on automated mapping generation using LLMs. The use case study (Chapter 6) evaluated six real-world scenarios across three mapping types. In all six scenarios the guided condition outperformed the baseline on semantic correctness. The most impactful fields were *mmv:hasDomainAssumption, mmv:hasAssumption, mmv:hasRiskOrIssue,* and *mmv:hasDesignDecision*, all from the Analysis and Design phases of LMMD.

7.2 Research Contributions

This section briefly revisits the contributions of this thesis, which were initially presented in Chapter 1.

The major contribution of this thesis is the design and development of LMMD, its formalisation as MMV, and its operationalisation through MetaSEMAP. LMMD advances the state of the art by providing the first lifecycle-based metadata model to apply across all three major types of declarative mappings. It is grounded in empirically validated requirements from 67 practitioners and evaluated across five

independent studies. LMMD complements existing approaches such as SSSOM and FAIR-IMPACT by organising metadata across lifecycle phases including analysis, design, development, testing, and maintenance and extending coverage to include testing and quality documentation that previous approaches did not emphasise. This lifecycle structure enables mapping artefacts to be understood, assessed, and reused.

The minor contribution of this thesis is a study on metadata-guided LLM mapping generation. The study demonstrated that lifecycle-based metadata consistently improves the semantic correctness of LLM-generated mappings across six real-world scenarios and three mapping types. The findings indicate that LMMD fields can function as structured generation templates, introducing a new use case for lifecycle-based metadata models in AI-assisted mapping engineering and demonstrating that metadata designed for documentation purposes can also guide automated generation

7.3 Impact and Uptake

The research described in this thesis has impacted the community through six publications spanning the full work from the initial proposal and community survey, through the ontology design and usability study, to the early evaluation of metadata-guided LLM generation. All publications received positive feedback from the research community. Venues include the IEEE International Conference on Semantic Computing (ICSC 2023, 2024, 2026), the SEMANTiCS 2025 conference and its co-located workshops (UKG 2025 and KG-NeSy 2025), and the ISWC Ontology Matching Workshop 2022.

MMV is publicly available under a CC-BY 4.0 licence at https://sarahalzahranitcd.github.io/MMV-Ontology2/ with Widoco-generated documentation. MetaSEMAP is publicly accessible at (https://metasemap-1-sarah-alzahrani.replit.app/), The research directly addresses part of the goals of ongoing international initiatives FAIR-IMPACT and the RDA FAIR Mappings Working

Group by providing a concrete, machine-readable vocabulary for lifecycle-aware mapping documentation. Further collaboration with these initiatives is anticipated.

7.4 Future Work

This section discusses potential future work for the research outlined in this thesis. The most significant direction for future work concerns the representation of purpose in semantic artefacts. As discussed in Section 5.5, the expert evaluation of MMV identified that string-valued properties such as hasPurpose and hasMappingDomain should be converted to object properties linking to named resources. Study 3 provided evidence that no standardised controlled vocabulary currently exists for representing the purpose of semantic artefacts a gap confirmed by the consensus of all six experts. This limitation extends beyond mapping projects to ontologies, datasets, and workflows of all kinds. The research community would benefit from a standardised purpose vocabulary for semantic artefacts.

Future work should also evaluate LMMD and MetaSEMAP with larger and more diverse participant groups, including practitioners from industry settings and domains outside the Semantic Web research community. Evaluating the lifecycle metadata model (LMMD) over the full lifecycle of a real mapping project would provide stronger evidence of practical adoption. Also, applying the model to real-world mapping projects where requirements evolve, data sources change, and multiple stakeholders are involved would test aspects of LMMD that controlled evaluation scenarios cannot fully address.

One of the challenges encountered during this research was the difficulty of finding and accessing mapping artefacts. Mappings are scattered across different communities, published in inconsistent naming conventions and formats, and accompanied by little or no structured metadata. Well-established initiatives such as OAEI publish alignment results annually through their websites, but without

associated structured metadata and without a shared repository. Integration of MMV with existing mapping development tools and emerging mapping repositories, would allow lifecycle metadata to be captured progressively rather than retrospectively. Such integration would enable researchers and practitioners to access mapping datasets that are annotated, validated, and described with documented quality metrics addressing a gap in the current mapping infrastructure.

Finally, the use case study presented in Chapter 6 evaluated a small number of scenarios with a single LLM.. The emerging research on LLM-based mapping generation increasingly reports that more context and semantics in the prompt lead to more complete and accurate outputs. Evaluating the effect of partial metadata and extending the LMMD-based prompt template across a wider range of models and mapping tasks are identified as directions for future work.

7.5 Final Remarks

It is hoped that LMMD, MMV, and MetaSEMAP will benefit the mapping community by providing a structured approach to documenting, reusing, and generating declarative mappings. Practitioners involved in the creation and maintenance of mappings can use the model to capture lifecycle context that is currently lost, making mapping artefacts easier to understand, assess, and reuse. It is also hoped that MMV can be adopted by tools and repositories involved in the mapping publication process to provide machine-readable lifecycle documentation alongside mapping artefacts.

Researchers working on mapping governance, quality, and AI-assisted generation can benefit from the findings and evaluation strategy presented in this thesis. Finally, it is hoped that the identification of a purpose vocabulary gap in Study 3 will inspire the research community to develop a shared standard for representing purpose across semantic artefacts, contributing to long-term improvements in mapping transparency and interoperability.

References

[1] A. Ferrara and A. Nikolov, “Data Linking for the Semantic Web,” *International Journal on Semantic Web and Information Systems*, vol. 7, no. 3, pp. 46–76, Jul. 2011, doi: 10.4018/jswis.2011070103.

[2] P. Shvaiko and J. Euzenat, “Ontology Matching: State of the Art and Future Challenges,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 1, pp. 158–176, 2013, doi: 10.1109/TKDE.2011.253.

[3] A. Dimou, M. V. Sande, P. Colpaert, E. Mannens, R. Verborgh, and R. V. de Walle, “RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data,” in *Proceedings of the 7th Workshop on Linked Data on the Web (LDOW)*, 2014. [Online]. Available: http://ceur-ws.org/Vol-1184/ldow2014_paper_01.pdf

[4] A. Dimou, T. De Nies, R. Verborgh, E. Mannens, P. Mechant, and R. Van de Walle, “Automated metadata generation for Linked Data generation and publishing workflows,” in *LDOW2016*, CEUR-WS. org, 2016, pp. 1–10.

[5] C. Debruyne, B. Walshe, and D. O’Sullivan, “Towards a Project Centric Metadata Model and Lifecycle for Ontology Mapping Governance,” in *Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services*, in iiWAS ’15, vol. 1–10. New York, NY, USA: Association for Computing Machinery, 2015, p. Article 50. doi: 10.1145/2837185.2837201.

[6] C. Zinke-Wehlmann *et al.*, “Linked Data and Metadata,” in *Big Data in Bioeconomy*, Springer, Cham, 2021, pp. 79–90.

[7] N. F. Noy and D. L. McGuinness, “Ontology Development 101: A Guide to Creating Your First Ontology,” p. 25.

[8] A. Randles, A. C. Junior, and D. O’Sullivan, “A Vocabulary for Describing Mapping Quality Assessment, Refinement and Validation,” in *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, 2021, pp. 425–430. doi: 10.1109/ICSC50631.2021.00076.

[9] Y. Le Franc *et al.*, “D4.5 — Guidelines and methodology to create, document and share mappings and crosswalks.,” FAIR-IMPACT, 2025. doi: 10.5281/zenodo.15111167.

[10] “DCMI Metadata Terms,” DCMI. Accessed: Jan. 12, 2026. [Online]. Available: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

[11] T. Lebo *et al.*, “PROV-O: The PROV Ontology,” *W3C Recommendation*, 2013, [Online]. Available: https://www.w3.org/TR/prov-o/

[12] W3C Data Exchange Working Group, “Data Catalog Vocabulary (DCAT) – Version 3,” World Wide Web Consortium (W3C), 2024. [Online]. Available: https://www.w3.org/TR/vocab-dcat-3/

[13] “Data on the Web Best Practices: Data Quality Vocabulary.” Accessed: Dec. 16, 2025. [Online]. Available: https://www.w3.org/TR/vocab-dqv/

[14] N. Matentzoglu *et al.*, “A Simple Standard for Sharing Ontological Mappings (SSSOM),” *Database*, vol. 2022, p. baac035, Jan. 2022, doi: 10.1093/database/baac035.
[15] P. Heyvaert, A. Dimou, and R. Verborgh, “YARRRML: Yet Another RML Rendering Language,” in *Proceedings of the 1st International Workshop on Knowledge Graph Construction (KGC 2018)*, in CEUR Workshop Proceedings, vol. 2160. CEUR-WS.org, 2018. [Online]. Available: http://ceur-ws.org/Vol-2160/paper_6.pdf
[16] “Ontology Alignment Evaluation Initiative::Home.” Accessed: Dec. 15, 2025. [Online]. Available: https://oaei.ontologymatching.org/
[17] “Metadata IG - RDA,” https://www.rd-alliance.org/. Accessed: Jan. 12, 2026. [Online]. Available: https://www.rd-alliance.org/groups/metadata-ig/activity/
[18] “FAIR Mappings WG - RDA,” https://www.rd-alliance.org/. Accessed: Jan. 12, 2026. [Online]. Available: https://www.rd-alliance.org/groups/fair-mappings-wg/activity/
[19] M. Hofer, J. Frey, and E. Rahm, “Towards self-configuring Knowledge Graph Construction Pipelines using LLMs – A Case Study with RML,” in *Proceedings of the 5th International Workshop on Knowledge Graph Construction (KGCW’24)*, 2024.
[20] D. Chaves-Fraga, F. Priyatna, A.-C. N. Ngomo, and J. D. Fernández, “Using Large Language Models for Semantic Data Integration Tasks: An Experimental Evaluation,” *arXiv preprint arXiv:2305.14367*, 2023, doi: 10.48550/arXiv.2305.14367.
[21] T. Berners-Lee, J. Hendler, and O. Lassila, “The Semantic Web" in Scientific American,” *Scientific American Magazine*, vol. 284, May 2001.
[22] “RDF 1.1 Concepts and Abstract Syntax.” Accessed: Apr. 04, 2026. [Online]. Available: https://www.w3.org/TR/rdf11-concepts/
[23] “OWL 2 Web Ontology Language Document Overview (Second Edition).” Accessed: Apr. 04, 2026. [Online]. Available: https://www.w3.org/TR/owl2-overview/
[24] A. Hogan *et al.*, “Knowledge Graphs,” *ACM Comput. Surv.*, vol. 54, no. 4, p. 71:1-71:37, Jul. 2021, doi: 10.1145/3447772.
[25] “R2RML: RDB to RDF Mapping Language.” Accessed: Apr. 04, 2026. [Online]. Available: https://www.w3.org/TR/r2rml/
[26] F. Scharffe and D. Fensel, “Correspondence Patterns for Ontology Alignment,” in *Knowledge Engineering: Practice and Patterns*, vol. 5268, A. Gangemi and J. Euzenat, Eds., in Lecture Notes in Computer Science, vol. 5268. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 83–92. doi: 10.1007/978-3-540-87696-0_10.
[27] F. Scharffe, “Correspondence Patterns Representation,” PhD Thesis, University of Innsbruck, 2009.
[28] B. Walshe, “Detecting Restriction Class Correspondences in Linked Open Data,” PhD Thesis, Trinity College Dublin, 2014.
[29] A. Iglesias-Molina *et al.*, “The RML Ontology: A Community-Driven Modular Redesign After a Decade of Experience in Mapping

Heterogeneous Data to RDF," in *The Semantic Web – ISWC 2023: 22nd International Semantic Web Conference, Athens, Greece, November 6–10, 2023, Proceedings, Part II*, Berlin, Heidelberg: Springer-Verlag, Nov. 2023, pp. 152–175. doi: 10.1007/978-3-031-47243-5_9.

[30] "RDF Mapping Language (RML)." Accessed: Mar. 11, 2026. [Online]. Available: https://rml.io/specs/rml/

[31] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov, "Silk – A Link Discovery Framework for the Web of Data," presented at the 2nd Workshop on Linked Data on the Web (LDOW 2009), 2009.

[32] A.-C. N. Ngomo and S. Auer, "LIMES - A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data," presented at the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011), 2011, pp. 2312–2317.

[33] "SPARQL 1.1 Query Language." Accessed: Dec. 11, 2023. [Online]. Available: https://www.w3.org/TR/sparql11-query/

[34] J. Euzenat, "An API for Ontology Alignment," in *The Semantic Web – ISWC 2004*, S. A. McIlraith, D. Plexousakis, and F. van Harmelen, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 698–712.

[35] J. David, J. Euzenat, F. Scharffe, and C. Trojahn dos Santos, "The Alignment API 4.0," *Semantic Web*, vol. 2, no. 1, pp. 3–10, Feb. 2011, doi: 10.3233/SW-2011-0028.

[36] H. Thomas, R. Brennan, and D. O'Sullivan, "MooM – A Prototype Framework for Management of Ontology Mappings," in *Proceedings - International Conference on Advanced Information Networking and Applications, AINA*, Mar. 2011, pp. 548–555. doi: 10.1109/AINA.2011.55.

[37] A. Mishra, "A Comparative Study of Different Software Development Life Cycle Models in Different Scenarios," *International Journal of Advance Research in Computer Science and Management Studies*, vol. 1, pp. 64–69, Oct. 2013.

[38] D. O'Sullivan, V. Wade, and D. Lewis, "Understanding as We Roam," *Internet Computing, IEEE*, vol. 11, pp. 26–33, Apr. 2007, doi: 10.1109/MIC.2007.50.

[39] H. Thomas, R. Brennan, and D. O'Sullivan, "Using the OM2R Meta-Data Model for Ontology Mapping Reuse for the Ontology Alignment Challenge: A Case Study," in *Proceedings of the 7th International Conference on Ontology Matching - Volume 946*, in OM'12. Aachen, DEU: CEUR-WS.org, 2012, pp. 61–72.

[40] J. Martínková, N. Juty, C. Goble, and A. Gonzalez Beltran GonzalezGoble, "Moving towards FAIR mappings and crosswalks," in *Proceedings of FOAM@FOIS 2024*, CEUR-WS.org, 2024. [Online]. Available: https://ceur-ws.org/Vol-3882/

[41] F. Orlandi, D. Graux, and D. O'Sullivan, "Benchmarking RDF Metadata Representations: Reification, Singleton Property and RDF," in *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, 2021, pp. 233–240. doi: 10.1109/ICSC50631.2021.00049.

[42] V. Nguyen, O. Bodenreider, and A. Sheth, “Don’t like RDF Reification? Making Statements about Statements Using Singleton Property,” in *Proceedings of the 23rd International Conference on World Wide Web*, in WWW ’14. New York, NY, USA: Association for Computing Machinery, 2014, pp. 759–770. doi: 10.1145/2566486.2567973.
[43] “RDF-star and SPARQL-star.” Accessed: Jun. 27, 2022. [Online]. Available: https://www.w3.org/2021/12/rdf-star.html
[44] O. Hartig, “Foundations of RDF* and SPARQL*:(An alternative approach to statement-level metadata in RDF),” in *AMW 2017 11th Alberto Mendelzon International Workshop on Foundations of Data Management and the Web, Montevideo, Uruguay, June 7-9, 2017.*, Juan Reutter, Divesh Srivastava, 2017.
[45] J. J. Carroll, C. Bizer, P. Hayes, and P. Stickler, “Named Graphs, Provenance and Trust,” in *Proceedings of the 14th International Conference on World Wide Web*, in WWW ’05. New York, NY, USA: Association for Computing Machinery, 2005, pp. 613–622. doi: 10.1145/1060745.1060835.
[46] D. Hernández, A. Hogan, and M. Krötzsch, “Reifying RDF: What Works Well With Wikidata?,” in *SSWS@ISWC*, T. Liebig and A. Fokoue, Eds., in CEUR Workshop Proceedings, vol. 1457. CEUR-WS.org, 2015, pp. 32–47. [Online]. Available: http://dblp.uni-trier.de/db/conf/semweb/ssws2015.html#HernandezHK15
[47] M. D. Wilkinson *et al.*, “The FAIR Guiding Principles for scientific data management and stewardship,” *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.
[48] “FAIR-IMPACT.” Accessed: Apr. 06, 2026. [Online]. Available: https://fair-impact.eu/
[49] N. Matentzoglu *et al.*, “A simple standard for ontological mappings 2023: updates on data model, collaborations and tooling.,” in *OM@ ISWC*, 2023, pp. 73–78.
[50] J. Toledo, A. Iglesias-Molina, D. Chaves-Fraga, and D. Garijo, “RMLdoc: Documenting Mapping Rules for Knowledge Graph Construction,” in *The Semantic Web: ESWC 2024 Satellite Events*, vol. 15344, A. Meroño Peñuela, O. Corcho, P. Groth, E. Simperl, V. Tamma, A. G. Nuzzolese, M. Poveda-Villalón, M. Sabou, V. Presutti, I. Celino, A. Revenko, J. Raad, B. Sartini, and P. Lisena, Eds., in Lecture Notes in Computer Science, vol. 15344. , Cham: Springer Nature Switzerland, 2025, pp. 326–330. doi: 10.1007/978-3-031-78952-6_51.
[51] A. Iglesias-Molina *et al.*, “The RML Ontology: A Community-Driven Modular Redesign After a Decade of Experience in Mapping Heterogeneous Data to RDF,” in *The Semantic Web – ISWC 2023*, T. R. Payne, V. Presutti, G. Qi, M. Poveda-Villalón, G. Stoilos, L. Hollink, Z. Kaoudi, G. Cheng, and J. Li, Eds., Cham: Springer Nature Switzerland, 2023, pp. 152–175.
[52] “FAIRCORE4EOSC.” Accessed: Apr. 06, 2026. [Online]. Available: https://faircore4eosc.eu/
[53] V. Emonet, J. Bolleman, S. Duvaud, T. M. de Farias, and A. C. Sima, “LLM-based SPARQL Query Generation from Natural Language over

Federated Knowledge Graphs," Feb. 10, 2025, *arXiv*: arXiv:2410.06062. doi: 10.48550/arXiv.2410.06062.
[54] H. Babaei Giglou, J. D'Souza, F. Engel, and S. Auer, "LLMs4OM: Matching Ontologies with Large Language Models," in *The Semantic Web: ESWC 2024 Satellite Events*, vol. 15344, A. Meroño Peñuela, O. Corcho, P. Groth, E. Simperl, V. Tamma, A. G. Nuzzolese, M. Poveda-Villalón, M. Sabou, V. Presutti, I. Celino, A. Revenko, J. Raad, B. Sartini, and P. Lisena, Eds., in Lecture Notes in Computer Science, vol. 15344. , Cham: Springer Nature Switzerland, 2025, pp. 25–35. doi: 10.1007/978-3-031-78952-6_3.
[55] W. J. Schmidt, I. Grangel-González, A. Paschke, and E. Kharlamov, "LLM-Supported Manufacturing Mapping Generation," *Transactions on Graph Data and Knowledge (TGDK)*, vol. 3, no. 3, p. 5:1-5:22, 2025, doi: 10.4230/TGDK.3.3.5.
[56] S. Ashiddiiqi, F. J. Ekaputra, M. Z. C. Candra, and G. A. P. S. Soekidjo, "Reducing Human Effort in Ontology-Based Data Integration: An LLM-Assisted Mapping Pipeline," presented at the Seventh International Workshop on Knowledge Graph Construction@ESWC2026, Mar. 2026. Accessed: Apr. 10, 2026. [Online]. Available: https://openreview.net/forum?id=qjJF4sIRKj
[57] A. Randles and D. O'Sullivan, "R2[RML]-ChatGPT Framework".
[58] J. Frey, L.-P. Meyer, F. Brei, S. Gründer-Fahrer, and M. Martin, "Assessing the Evolution of LLM Capabilities for Knowledge Graph Engineering in 2023," in *The Semantic Web: ESWC 2024 Satellite Events*, A. Meroño Peñuela, O. Corcho, P. Groth, E. Simperl, V. Tamma, A. G. Nuzzolese, M. Poveda-Villalón, M. Sabou, V. Presutti, I. Celino, A. Revenko, J. Raad, B. Sartini, and P. Lisena, Eds., Cham: Springer Nature Switzerland, 2025, pp. 51–60. doi: 10.1007/978-3-031-78952-6_5.
[59] S. Lushnei, D. Shumskyi, S. Shykula, and E. Jimenez-Ruiz, "Large Language Models as Oracles for Ontology Alignment".
[60] M. Taboada, D. Martinez, M. Arideh, and R. Mosquera, "Ontology Matching with Large Language Models and Prioritized Depth-First Search," arXiv.org. Accessed: Apr. 11, 2026. [Online]. Available: https://arxiv.org/abs/2501.11441v2
[61] R. Ihtassine, A. F. Firmansyah, N. Srivastava, M. Ali, A.-C. Ngonga Ngomo, and M. A. Sherif, "NL2LS: LLM-based Automatic Linking of Knowledge Graphs," in *Proceedings of the 13th Knowledge Capture Conference 2025*, in K-CAP '25. New York, NY, USA: Association for Computing Machinery, Dec. 2025, pp. 227–231. doi: 10.1145/3731443.3771374.
[62] "Large Language Models for Ontology Engineering: A Systematic Literature Review | www.semantic-web-journal.net." Accessed: Apr. 11, 2026. [Online]. Available: https://www.semantic-web-journal.net/content/large-language-models-ontology-engineering-systematic-literature-review
[63] G. Agrawal, T. Kumarage, Z. Alghamdi, and H. Liu, "Can Knowledge Graphs Reduce Hallucinations in LLMs? : A Survey," in *Proceedings of the*

2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), K. Duh, H. Gomez, and S. Bethard, Eds., Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 3947–3960. doi: 10.18653/v1/2024.naacl-long.219.

[64] M. Freund, R. Dorsch, S. Schmid, and A. Harth, "Mapping by Example: Towards an RML Mapping Reverse Engineering Pipeline," in *ESWC 2025 Workshop on Knowledge Graph Construction (KGCW)*, 2025.

[65] S. Alzahrani and D. O'Sullivan, "What metadata is needed for semantic and data mappings?," in *2023 IEEE 17th International Conference on Semantic Computing (ICSC)*, 2023, pp. 312–316. doi: 10.1109/ICSC56153.2023.00064.

[66] S. Alzahrani and D. Sullivan, "Towards a Unified Metadata Model for Semantic and Data Mappings," in *Ontology Matching workshop at ISWC 2022*, 2022.

[67] M. Poveda-Villalón, A. Fernández-Izquierdo, and R. García-Castro, "Linked Open Terms (LOT) Methodology (Version 1.0)." Ontology Engineering Group, Universidad Politécnica de Madrid, 2019. doi: 10.5281/zenodo.2539305.

[68] "The NeOn Methodology for Ontology Engineering | Springer Nature Link." Accessed: Apr. 07, 2026. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-24794-1_2

[69] S. Chávez-Feria, R. García-Castro, and M. Poveda-Villalón, "Chowlk: from UML-Based Ontology Conceptualizations to OWL," in *The Semantic Web*, vol. 13261, P. Groth, M.-E. Vidal, F. Suchanek, P. Szekley, P. Kapanipathi, C. Pesquita, H. Skaf-Molli, and M. Tamper, Eds., in Lecture Notes in Computer Science, vol. 13261. , Cham: Springer International Publishing, 2022, pp. 338–352. doi: 10.1007/978-3-031-06981-9_20.

[70] S. Alzahrani and D. O'Sullivan, "MetaMap: Facilitating Mapping Reuse and Lifecycle Documentation," in *2024 IEEE 18th International Conference on Semantic Computing (ICSC)*, 2024, pp. 323–328. doi: 10.1109/ICSC59802.2024.00058.

[71] S. Alzahrani and D. O'Sullivan, "MetaSEMAP: Usability Evaluation for Metadata Annotation Across the Mapping Lifecycle," in *First International Workshop on Users and Knowledge Graphs@ SEMANTiCS 2025*,

[72] S. Alzahrani and D. O'Sullivan, "Guiding LLM Generated Mappings with Lifecycle-Based Metadata: An Early Evaluation," in *Joint Proceedings of Posters, Demos, Workshops, and Tutorials of the 21st International Conference on Semantic Systems co-located with 21st International Conference on Semantic Systems (SEMANTiCS 2025), Vienna, Austria, September 3-5, 2025*, D. Chaves-Fraga, I. Heibi, D. Garijo, D. Collarana, A. A. Salatino, and S. Vahdati, Eds., in CEUR Workshop Proceedings, vol. 4064. CEUR-WS.org, 2025. [Online]. Available: https://ceur-ws.org/Vol-4064/kgnesy-paper4.pdf

[73] B. A. Kitchenham and S. L. Pfleeger, “Principles of survey research part 2: designing a survey,” *SIGSOFT Softw. Eng. Notes*, vol. 27, no. 1, pp. 18–20, Jan. 2002, doi: 10.1145/566493.566495.
[74] L. J. Cronbach, “Coefficient alpha and the internal structure of tests,” *Psychometrika*, vol. 16, no. 3, pp. 297–334, Sep. 1951, doi: 10.1007/BF02310555.
[75] “Usability Engineering,” ScienceDirect. Accessed: Apr. 19, 2026. [Online]. Available: http://www.sciencedirect.com:5070/book/monograph/9780125184069/usability-engineering
[76] “Psychometric Evaluation of the PSSUQ Using Data from Five Years of Usability Studies: International Journal of Human–Computer Interaction: Vol 14, No 3-4.” Accessed: Apr. 19, 2026. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/10447318.2002.9669130
[77] john Brooke, “SUS: A ‘Quick and Dirty’ Usability Scale,” in *Usability Evaluation In Industry*, CRC Press, 1996.
[78] V. Braun and V. Clarke, “Using thematic analysis in psychology,” *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, Jan. 2006, doi: 10.1191/1478088706qp063oa.
[79] A. Randles and D. O’Sullivan, “Phased Evaluation of Ontologies with Domain and Ontology Experts,” in *15th Workshop on Ontology Design and Patterns*, 2024. [Online]. Available: https://openreview.net/forum?id=iZ17DHWe7v
[80] M. Poveda-Villalón, A. Gómez-Pérez, and M. C. Suárez-Figueroa, “OOPS! (OntOlogy Pitfall Scanner!): An On-line Tool for Ontology Evaluation,” *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 10, no. 2, pp. 7–34, 2014, doi: 10.4018/ijswis.2014040102.
[81] D. Garijo, O. Corcho, and M. Poveda-Villalon, “FOOPS!: An Ontology Pitfall Scanner for the FAIR Principles? ??”.
[82] S. Alzahrani and D. O'Sullivan, "MMV: A Lifecycle Based Ontology for Declarative Mapping Projects," in Proceedings of the IEEE 20th International Conference on Semantic Computing (ICSC 2026), in press.

Appendices

Label	Title	Page
Appendix A	Full Metadata Field Specifications	235
Appendix B	CQs, MMV Properties and SPARQL Query Templates	241
Appendix C	MMV new classes and properties tables	243
Appendix D	Study 1 Materials	245
Appendix E	Study 2 Materials	261
Appendix F	Study 3 Materials	267
Appendix G	Study 4 Materials	277
Appendix H	Use Case (Chapter 6) Materials	298
Appendix I	Published papers	320

Appendix A : Full Metadata Field Specifications

This appendix provides the full metadata field specifications for the metadata models reviewed in Chapter 3, including lifecycle-based models (Section 3.2.1) and artefact-centric approaches (Section 3.2.2)

A.1: OM2R Metadata Model Full Field Specifications [39]

The following table presents the complete set of metadata fields implemented in the OM2R ontology as described in Thomas et al. [35]. Field names use the original ontology property names as implemented in OWL-DL.

Phase	Field Name	Type	Description
Characterisation	om2r:human_readable_name	String	Human-readable name of source and target ontologies
	om2r:description	String	Description of source and target ontologies
	om2r:hasLocation	URL	Location where ontology can be accessed or downloaded
	om2r:hasCreationDate	Date	Creation date of ontology
	om2r:hasIdentifier	URI	Unique identifier for the ontology
	om2r:hasVersion	URI	Version of the ontology
	om2r:hasClassCount	Integer	Number of classes in the ontology
	om2r:hasInstanceCount	Integer	Number of instances in the ontology
	om2r:hasPropertyClass	Integer	Number of properties in the ontology
	om2r:hasDesign	Controlled vocabulary	Basic design classification (e.g. deep_hierarchy, flat_hierarchy)
	om2r:hasFormalLanguage	Controlled vocabulary	Formal language used (e.g. OWL-DL, RDF-S)
	om2r:hasNotation	Controlled vocabulary	Exchange notation (e.g. RDF/XML)
	om2r:Editor	String	Editor or contact information
Matching	Matching:hasLocation	URL	Location of alignment file
	om2r:hasFormalMatchingLanguage	Controlled vocabulary	Formal language used for matching (e.g. EDOAL)
	om2r:hasNotation	Controlled vocabulary	Exchange notation of matching file (e.g. RDF/XML)
	om2r:hasMethod	Controlled vocabulary	Matching method (manual / automatic / mixed)
	om2r:hasMatchingTool	String	Tool used to generate alignment
	Algorithm:encodedIn	String	Programming language of algorithm implementation
	Algorithm:hasJavaClass	String	Java class name of matching algorithm implementation

	URL	URL	Source code location
	om2r:isBasedOn	Controlled vocabulary	Elements analysed (e.g. rdfs:label, rdfs:class)
	om2r:has_Applied_Threshold	Float	Similarity threshold value applied
	om2r:hasScope	Controlled vocabulary	Matching scope (complete / partial)
	om2r:hasMatchRequirements	String	Specific requirements for the matching task
Mapping/Execution	–	–	Confirmed mappings, rendering formats, storage references
Management	om2r:asVersionId	String	Version identifier
	om2r:hasHashID	String	Hash ID for automated change detection
	ChangeLog	String	Textual description of changes made
Metadata Creation	–	–	Parallel activity, metadata collected and processed across all phases

Specific field names for the Mapping/Execution phase are not fully specified in the original OM2R publication [35] beyond the general categories listed.

A.2: Project-Centric Governance Model Full Field Specifications [5]

The following table presents the complete set of concepts and relations implemented in the project-centric governance ontology as described in Debruyne et al [5]. Concept names use the original mg: prefix as implemented in OWL-2, extending PROV-O.

Phase	Field Name	Type	Description
Stage	mg:Stage	Activity	First activity, project inception and scoping
	mg:Requirement	Entity	Documented condition the alignment must satisfy
	mg:Scope	Entity	Features characterising the alignment and mapping
	mg:Community	Entity	Group of human agents contributing to the project
	prov:Person	Entity	Individual stakeholder within the community
	mg:withStakeholder	Property	Relates community to individual stakeholder roles

Characterise	mg:Characterize	Activity	Discovery and analysis of ontologies to be mapped
	mg:DiscoveryReport	Entity	Summary of ontology discovery process
	mg:OntologyAnalysis	Entity	Assessment of ontology amenability for mapping
	mg:Selection	Entity	Selected source and target ontologies
	mg:MappingDecision	Entity	Decision whether to proceed with the mapping project
Reuse	mg:Reuse	Activity	Discovery and assessment of reusable alignments
	mg:AlignmentRepository	Agent	Repository queried for reusable alignments
	mg:AlignmentSearchEngine	Agent	Search engine used to discover existing alignments
	mg:ReuseEvaluation	Entity	Evaluation report of candidate alignments for reuse
	mg:ReuseDecision	Entity	Decision , direct reuse / adaptation / new alignment
Match	mg:Plan	Activity	Planning of matcher configurations and execution
	mg:MatcherConfiguration	Entity	Combination of matcher + parameters + resources
	mg:Matcher	Agent	Matching tool or algorithm used
	mg:Execute	Activity	Execution of matching algorithms on ontologies
	mg:CandidateCorrespondences	Entity	Set of correspondence candidates generated
	mg:Evaluate	Activity	Evaluation of candidate correspondences
	mg:MatchingEvaluationDecision	Entity	Decision to proceed or return to planning
Align and Map	mg:Prepare	Activity	Planning the creation of the alignment
	mg:AlignmentPlan	Entity	Actions to be undertaken during alignment creation
	mg:Create	Activity	Creation of alignment from candidate correspondences
	mg:Alignment	Entity	Final set of validated correspondences
	mg:Render	Activity	Generation of executable mapping from alignment
	mg:Mapping	Entity	Executable mapping artefact
Application	mg:Application	Activity	Use of rendered mapping in target application
	FeedbackEvent	Entity	Event triggered by application insights or data changes
	OntologyVersionLog	Entity	Log monitoring ontology changes affecting existing mappings

Type values reflect PROV-O and OWL-2 ontology concept categories as implemented in Debruyne et al. [5]., where Activity, Entity, and Agent are subclasses of prov:Activity, prov:Entity, and prov:Agent respectively.

A.3: SSSOM Full Field Specifications [14], [49]

The following table presents the complete set of metadata fields implemented in SSSOM as described in Matentzoglu et al. [14]. Fields are organised by level: correspondence level and mapping set level, and the Required column indicates the status of each field as defined in the official SSSOM specification

Level	Field	Description
Correspondence Level	subject id	The identifier of the subject entity being mapped
	predicate id	The predicate or relation linking the subject and object
	object id	The identifier of the object entity being mapped
	mapping justification	Justification for the mapping using SEMAPV controlled terms (e.g. LexicalMatching, ManualMappingCuration)
	subject label	Human-readable label of the subject entity
	object label	Human-readable label of the object entity
	confidence	Creator's confidence that the mapping is correct (0.0–1.0)
	author id	Identifier of the person or group who asserted the mapping
	creator id	Identifier of the person or group who created the mapping file
	reviewer id	Identifier of the person or group who reviewed the mapping
	mapping date	Date the mapping was asserted
	mapping tool	Tool or algorithm used to generate the mapping
	subject source	URI of the vocabulary or source for the subject entity
	object source	URI of the vocabulary or source for the object entity
	subject source version	Version of the source vocabulary for the subject
	object source version	Version of the source vocabulary for the object
	subject match field	Properties used to establish the match on the subject side
	object match field	Properties used to establish the match on the object side
	match string	String shared by subject and object used for matching
	subject preprocessing	Preprocessing method applied to subject fields before matching
	object preprocessing	Preprocessing method applied to object fields before matching
	similarity score	Numerical score denoting similarity between entities (0–1)
	similarity measure	The measure used for computing the similarity score
	see also	URL providing additional context for the mapping instance
	comment	Free text curator notes or additional informative information
	other	Key-value pairs for properties not covered by the standard fields
Mapping Set Level	mapping set id	Globally unique identifier for the mapping set
	license	URL to the license of the mapping set
	mapping set version	Version of the mapping set
	mapping set title	Human-readable title of the mapping set
	mapping set description	Description of the purpose and scope of the mapping set
	creator id	Identifier of the creators of the mapping set
	mapping date	Date the mapping set was created or published
	mapping tool version	Version string of the mapping tool used

SSSOM has continued to evolve since its original publication, with additional fields introduced in subsequent community revisions [49].The fields listed here reflect those described in the original 2022 paper [14]

A.4: Alignment API Format Full Field Specifications [34] [35]

The following table presents the complete set of metadata fields defined in the Alignment API format and its EDOAL extension, as described in Euzenat [34] and David et al. [35] Fields are organised by level alignment level, correspondence level, and EDOAL extension.

Level	Field	Description
Alignment Level	Ontology 1	URI of the first ontology being aligned
	Ontology 2	URI of the second ontology being aligned
	type	Cardinality type of the alignment (11, 1n, n1, nn)
	level	Level of alignment (0 = URI only, 1 = labels, 2 = EDOAL)
	method	Method used to generate the alignment
	time	Time taken to generate the alignment
Correspondence Level	entity1	URI of the entity in the first ontology
	entity2	URI of the entity in the second ontology
	relation	Semantic relation between entities (=, <, >, %, !)
	measure	Numerical confidence value (0.0–1.0)
EDOAL Extension	Constructed entities	Entities defined through boolean or property construction operators (disjunction, conjunction, complement, inverse, composition)
	Property restrictions	Domain, range, cardinality, and value constraints on properties
	Property transformations	Unit conversions and value encoding transformations for property values
	Linkkeys	Conditions for establishing instance-level matches between ontologies

Note: The Alignment API format defines only 10 core fields across alignment and correspondence levels. The EDOAL extension adds expressive power for complex correspondences but does not add documentation or governance metadata such as author, date, license, or version.

A.5: RML Approaches, Full Field Specifications [4] [50], [51]

The following table presents the complete metadata fields and components captured by three RML documentation and formalisation approaches: the automated metadata generation approach by Dimou et al. [4], the RMLdoc documentation tool by Toledo et al. [50], and the RML Ontology by Iglesias-Molina et al. [51].

Approach	Category	Field / Component
Dimou et al.	Provenance (PROV-O)	Activity (mapping definition, data retrieval, RDF generation, publication), Entity (mapping document, raw data, RDF dataset), Agent (mapping editor, data generator, data publisher, data owner), timestamps (startedAtTime, endedAtTime), derivation relations (wasGeneratedBy, wasAttributedTo, used, wasAssociatedWith)
	Dataset (VoID)	Access URL, data dump location, number of triples, number of distinct subjects, number of distinct objects
	Dataset (DCAT)	Identifier, title, description, publisher, creator, issued date, modified date, access URL, download URL, format, licence
	Mapping-related	Mapping document reference, Triples Map associations, mapping editor agent, mapping definition activity timestamp
RMLdoc	Document level	Title, version, authors, file name, description, creation date, license
	Structural components	Logical source (data source reference), subject description, predicate-object descriptions, join conditions (references between rule sets), namespace prefixes
RML Ontology	Core mapping components	rml:LogicalSource, rml:SubjectMap, rml:PredicateObjectMap, rml:PredicateMap, rml:ObjectMap, rml:JoinCondition, rml:ReferenceFormulation, rml:TermMap, rml:GraphMap, rml:LanguageMap, rml:DatatypeMap, rml:TriplesMap

Note: Dimou et al. do not define a fixed set of named metadata fields. Instead, metadata is generated dynamically using established vocabularies PROV-O, VoID, and DCAT. Fields listed above reflect the metadata categories described in the paper. The RML Ontology focuses on structural components of mapping rules rather than documentation or lifecycle metadata.

Appendix B: CQs, MMV Properties and SPARQL Query Templates

This appendix presents the complete set of competency questions (CQs) used to guide the design of the Mapping Metadata Vocabulary (MMV). For each competency question, the table provides: the relevant MMV classes and properties that enable the question to be answered, and an example SPARQL query template.

CQ	Competency Question	MMV Classes / Properties	Example SPARQL Query Template
CQ1	What is the purpose and scope of the mapping project?	mmv:MappingProject mmv:AnalysisActivity mmv:hasPurpose mmv:hasRequirement mmv:hasMappingDomain	SELECT ?project ?purpose ?domain WHERE { ?project a mmv:MappingProject ; mmv:hasAnalysis ?analysis . ?analysis mmv:hasPurpose ?purpose . OPTIONAL { ?analysis mmv:hasMappingDomain ?domain } }
CQ2	Which mapping type is being documented?	mmv:MappingArtifact mqv:hasMappingType	SELECT ?artifact ?type WHERE { ?artifact a mmv:MappingArtifact ; mqv:hasMappingType ?type . }
CQ3	What assumptions or requirements does the mapping rely on?	mmv:AnalysisActivity mmv:hasAssumption mmv:hasDomainAssumption mmv:hasRequirement mmv:hasTechnicalRequirement mmv:hasRiskOrIssue	SELECT ?project ?assumption ?req WHERE { ?project mmv:hasAnalysis ?analysis . OPTIONAL { ?analysis mmv:hasAssumption ?assumption } OPTIONAL { ?analysis mmv:hasRequirement ?req } OPTIONAL { ?analysis mmv:hasDomainAssumption ?da } }
CQ4	Which lifecycle activities are associated with the mapping project?	mmv:MappingProject mmv:hasAnalysis mmv:hasDesign mmv:hasDevelopment mmv:hasTesting mmv:hasMaintenance	SELECT ?project ?activity WHERE { ?project a mmv:MappingProject . { ?project mmv:hasAnalysis ?activity } UNION { ?project mmv:hasDesign ?activity } UNION { ?project mmv:hasDevelopment ?activity } UNION { ?project mmv:hasTesting ?activity } UNION { ?project mmv:hasMaintenance ?activity } }
CQ5	What key design decisions were made and why?	mmv:DesignActivity mmv:hasDesignDecision mmv:hasJustification mmv:hasPlannedQualityMetric	SELECT ?project ?decision ?justification WHERE { ?project mmv:hasDesign ?design . ?design mmv:hasDesignDecision ?decision . OPTIONAL { ?design mmv:hasJustification ?justification } }
CQ6	Which mapping artefacts were produced and how are they described?	mmv:MappingArtifact mmv:hasMappingArtifact mmv:version mmv:hasDescription mqv:hasMappingType mqv:hasMappingFormat	SELECT ?project ?artifact ?type ?format ?version WHERE { ?project mmv:hasDevelopment ?dev . ?dev mmv:hasMappingArtifact ?artifact . OPTIONAL { ?artifact mqv:hasMappingType ?type } OPTIONAL { ?artifact mqv:hasMappingFormat ?format }

			OPTIONAL { ?artifact mmv:version ?version } }
CQ7	Which mapping language and format were used?	mmv:DevelopmentActivity mqv:hasMappingFormat mqv:hasMappingMethod	SELECT ?project ?format ?method WHERE { ?project mmv:hasDevelopment ?dev . OPTIONAL { ?dev mqv:hasMappingFormat ?format } OPTIONAL { ?dev mqv:hasMappingMethod ?method } }
CQ8	Which tools or methods were used during development?	mmv:DevelopmentActivity mmv:hasTool mmv:hasMappingAlgorithm	SELECT ?project ?tool ?algorithm WHERE { ?project mmv:hasDevelopment ?dev . OPTIONAL { ?dev mmv:hasTool ?tool } OPTIONAL { ?dev mmv:hasMappingAlgorithm ?algorithm } }
CQ9	How was the mapping tested and what were the results?	mmv:TestingActivity mmv:hasTestingType mmv:hasTestingResult mmv:QualityAssessment mmv:hasQualityMeasurement dqv:QualityMeasurement	SELECT ?project ?testType ?result WHERE { ?project mmv:hasTesting ?testing . OPTIONAL { ?testing mmv:hasTestingType ?testType } OPTIONAL { ?testing mmv:hasTestingResult ?result } OPTIONAL { ?testing mmv:hasQualityMeasurement ?qm . ?qm dqv:value ?score } }
CQ10	How is the mapping maintained and versioned over time?	mmv:MaintenanceActivity mmv:hasPublisherName mmv:hasPublisherSource mmv:version mmv:hasVersionDateTime	SELECT ?project ?version ?date ?publisher WHERE { ?project mmv:hasMaintenance ?maint . OPTIONAL { ?maint mmv:version ?version } OPTIONAL { ?maint mmv:hasVersionDateTime ?date } OPTIONAL { ?maint mmv:hasPublisherName ?publisher } }

Note: SPARQL templates are illustrative and assume MMV-compliant RDF instances.

Appendix C: MMV new classes and properties tables

Note: Only new MMV-specific classes and properties are listed in this appendix. Classes and properties reused from existing vocabularies including PROV-O, Dublin Core Terms, FOAF, DQV, MQV, and OWL are not listed here. The complete ontology documentation including all reused and new terms is available at https://sarahalzahranitcd.github.io/MMV-Ontology2/

Classes:

MMV Term	Description	Why existing vocabulary insufficient
mmv:MappingProject	Represents a declarative mapping effort as a first-class entity, acting as the contextual anchor for all lifecycle activities and artefacts.	PROV-O prov:Activity and prov:Entity are too generic neither captures a mapping project as a structured, lifecycle-bearing unit.
mmv:AnalysisActivity	Represents the analysis phase, capturing stakeholders, input datasets, requirements, assumptions, and risks before design.	prov:Activity cannot distinguish analysis phase activities from design or development activities.
mmv:DesignActivity	Represents the design phase, capturing modelling decisions and their justifications.	No existing vocabulary distinguishes design phase activities or captures design rationale in a mapping context.
mmv:DevelopmentActivity	Represents the implementation phase, capturing mapping language, tools, methods, and algorithms.	MQV captures some mapping metadata but does not model it as a lifecycle activity linked to a project.
mmv:TestingActivity	Represents validation and quality assessment activities, including test type, results, and quality measurements.	DQV provides quality metrics but does not model testing as a distinct lifecycle activity linked to a mapping project.
mmv:MaintenanceActivity	Represents post deployment activities including versioning, publication, and long-term governance.	No existing vocabulary models maintenance as a lifecycle activity with versioning and publication metadata.
mmv:MappingArtifact	Represents a concrete mapping specification (RML mapping, ontology alignment, or interlinking specification) as a named, versioned entity.	MQV mqv:Mapping focuses on quality aspects; it does not model the artefact as a lifecycle entity linked to design and testing activities.
mmv:QualityAssessment	Represents a quality assessment activity associated with a mapping artefact, linking it to metrics and measurements.	DQV dqv:QualityMeasurement captures measurements but does not model the assessment activity linked to a specific mapping project.
mmv:DataSet	Represents an input dataset used in the mapping project, described with source, format, type, and creator information.	DCAT dcat:Dataset is designed for data cataloguing; it does not capture the role of an input dataset within a mapping lifecycle.

Properties:

MMV Term	Description	Why existing vocabulary insufficient
mmv:hasAnalysis	Links a MappingProject to its AnalysisActivity instance.	No existing property links a project specifically to an analysis phase activity.
mmv:hasDesign	Links a MappingProject to its DesignActivity instance.	No existing property links a project specifically to a design phase.
mmv:hasDevelopment	Links a MappingProject to its DevelopmentActivity instance.	No existing vocabulary provides lifecycle-phase-specific linking properties.
mmv:hasTesting	Links a MappingProject to its TestingActivity instance.	No existing vocabulary provides lifecycle-phase-specific linking properties.
mmv:hasMaintenance	Links a MappingProject to its MaintenanceActivity instance.	No existing vocabulary provides lifecycle-phase-specific linking properties.
mmv:hasStakeholder	Links a MappingProject to a stakeholder agent.	prov:wasAttributedTo links entities to agents but does not capture stakeholder roles within a mapping project.
mmv:hasInput	Links an AnalysisActivity to a DataSet representing an input.	No existing vocabulary links an analysis activity to its input datasets within a mapping lifecycle.
mmv:hasMappingArtifact	Links a lifecycle activity to the mapping artefact it produced or tested.	prov:generated links activities to entities but cannot distinguish the specific role of a mapping artefact.
mmv:hasPlannedQualityMetric	Links a DesignActivity to quality metrics planned for the mapping.	DQV provides no property linking planned quality metrics to a design-phase activity.
mmv:hasQualityMeasurement	Links a TestingActivity to DQV quality measurement instances.	DQV provides no property linking measurements to a testing activity within a mapping lifecycle.
mmv:isAnalysisOf	Inverse of mmv:hasAnalysis links an AnalysisActivity back to its MappingProject.	Supports bidirectional traversal not available with generic PROV-O properties.
mmv:isDesignOf	Inverse of mmv:hasDesign.	Supports bidirectional traversal not available with generic PROV-O properties.
mmv:isDevelopmentOf	Inverse of mmv:hasDevelopment.	Supports bidirectional traversal not available with generic PROV-O properties.
mmv:isTestingOf	Inverse of mmv:hasTesting.	Supports bidirectional traversal not available with generic PROV-O properties.
mmv:isMaintenanceOf	Inverse of mmv:hasMaintenance.	Supports bidirectional traversal not available with generic PROV-O properties.
mmv:isMappingArtifactOf	Inverse of mmv:hasMappingArtifact.	Supports bidirectional traversal not available with generic PROV-O properties.

Appendix D: Study 1 Materials

D.1: Ethics application approval letter



Sarah Alzahrani <salzahra@tcd.ie>

TCD REC WebApp: The status of 'A Survey on Metadata for Semantic and Data Mappings' (1005) has been updated by the Committee

rec-app-help@tchpc.tcd.ie <rec-app-help@tchpc.tcd.ie> 9 February 2022 at 14:47
To: salzahra@tcd.ie

The status of 'A Survey on Metadata for Semantic and Data Mappings' has been updated by the Committee.

Title: 'A Survey on Metadata for Semantic and Data Mappings'
Applicant Name: Sarah Alzahrani
Submitted by: Sarah Alzahrani
Academic Supervisor: Declan O'Sullivan
Application Number: 20211203

Result of the REC Meeting: Approved

The Feedback from the Committee is as follows:
Dear Sarah,
This application has been approved and you can now complete your study.
Kind regards
Elaine

The application can be viewed here:

https://webhost.tchpc.tcd.ie/research_ethics/?q=node/1005

If amendments are required, please use the following link to edit the application and upload the changes:

https://webhost.tchpc.tcd.ie/research_ethics/?q=node/1005/edit

D.2: Invitation Email

Survey on metadata for semantics and data mappings

Sarah Alzahrani <sarah.alzahrani@adaptcentre.ie> Wed, Mar 9, 2022 at 11:00 AM

Dear .

I hope you are doing well,

As you might know, I am a first-year Ph.D. student under the supervision of Prof. Declan O'Sullivan. My research focuses on improving metadata and its representation for uplift, semantic, and interlink mappings, as well as reasoning over metadata for purposes such as discovery, reuse, and quality assessment.

I would like to invite you to participate in this study on metadata for semantic and data mapping. This study aims to identify a set of metadata fields that can be used to annotate different types of mappings. Your participation is completely voluntary and all of your responses are anonymous. None of the responses will be connected to identifying information. It is expected that your participation in this survey will not exceed 15 minutes.

To participate, please click on the following link: [survey link]

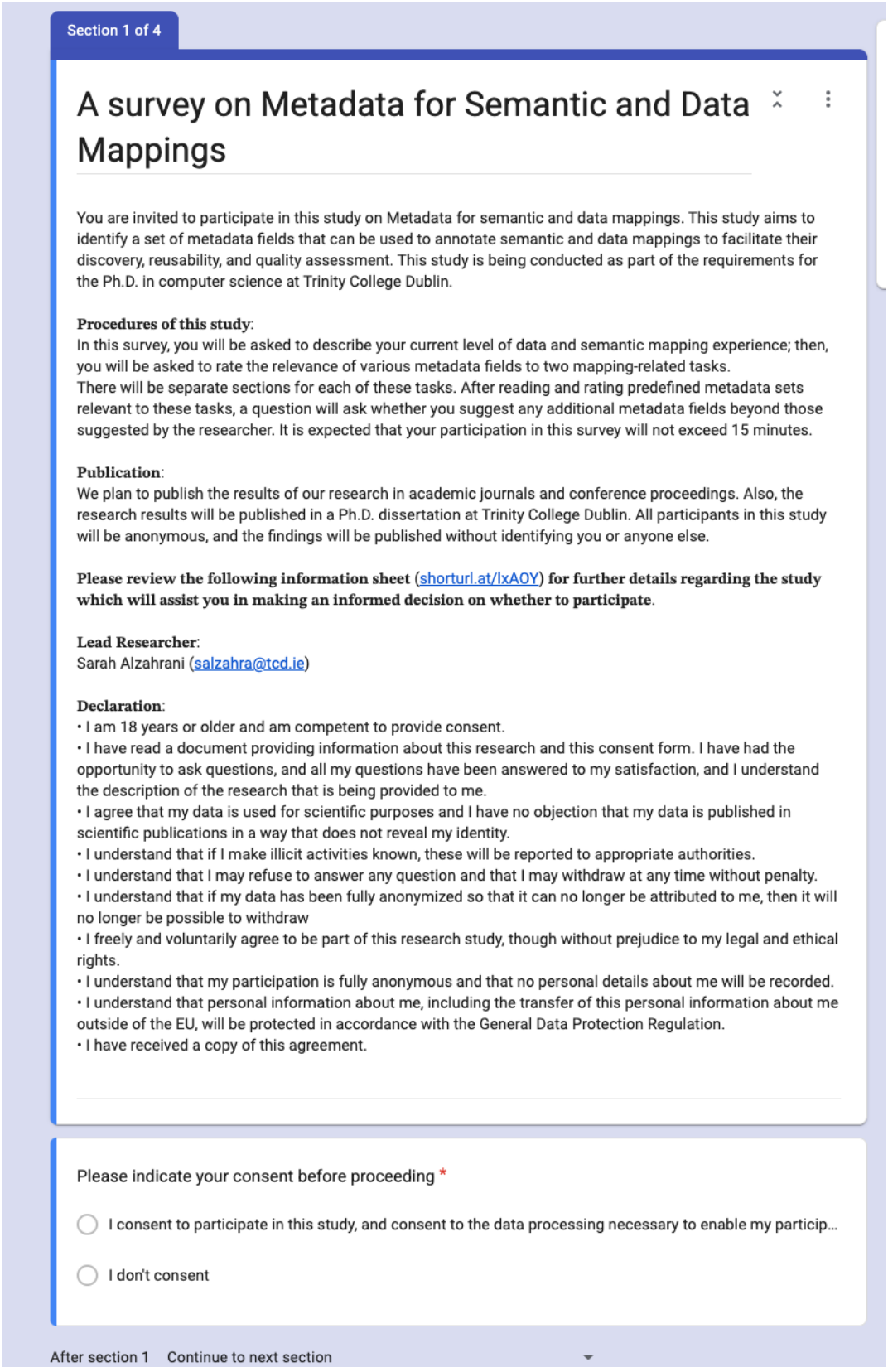
If you have any questions about this survey, or difficulty in accessing the site or completing the survey, please contact me at (salzahra@tcd.ie)

Your cooperation is highly appreciated and due to being extremely tight on time, it would be really helpful if you could complete the survey by the 18th of March.

Note: This survey has been approved according to the School of Computer Science and Statistics ethics policy.

Many Thanks,
Sarah

D.3: Consent Form



Section 1 of 4

A survey on Metadata for Semantic and Data Mappings

You are invited to participate in this study on Metadata for semantic and data mappings. This study aims to identify a set of metadata fields that can be used to annotate semantic and data mappings to facilitate their discovery, reusability, and quality assessment. This study is being conducted as part of the requirements for the Ph.D. in computer science at Trinity College Dublin.

Procedures of this study:
In this survey, you will be asked to describe your current level of data and semantic mapping experience; then, you will be asked to rate the relevance of various metadata fields to two mapping-related tasks.
There will be separate sections for each of these tasks. After reading and rating predefined metadata sets relevant to these tasks, a question will ask whether you suggest any additional metadata fields beyond those suggested by the researcher. It is expected that your participation in this survey will not exceed 15 minutes.

Publication:
We plan to publish the results of our research in academic journals and conference proceedings. Also, the research results will be published in a Ph.D. dissertation at Trinity College Dublin. All participants in this study will be anonymous, and the findings will be published without identifying you or anyone else.

Please review the following information sheet (shorturl.at/lxAOY) for further details regarding the study which will assist you in making an informed decision on whether to participate.

Lead Researcher:
Sarah Alzahrani (salzahra@tcd.ie)

Declaration:
• I am 18 years or older and am competent to provide consent.
• I have read a document providing information about this research and this consent form. I have had the opportunity to ask questions, and all my questions have been answered to my satisfaction, and I understand the description of the research that is being provided to me.
• I agree that my data is used for scientific purposes and I have no objection that my data is published in scientific publications in a way that does not reveal my identity.
• I understand that if I make illicit activities known, these will be reported to appropriate authorities.
• I understand that I may refuse to answer any question and that I may withdraw at any time without penalty.
• I understand that if my data has been fully anonymized so that it can no longer be attributed to me, then it will no longer be possible to withdraw
• I freely and voluntarily agree to be part of this research study, though without prejudice to my legal and ethical rights.
• I understand that my participation is fully anonymous and that no personal details about me will be recorded.
• I understand that personal information about me, including the transfer of this personal information about me outside of the EU, will be protected in accordance with the General Data Protection Regulation.
• I have received a copy of this agreement.

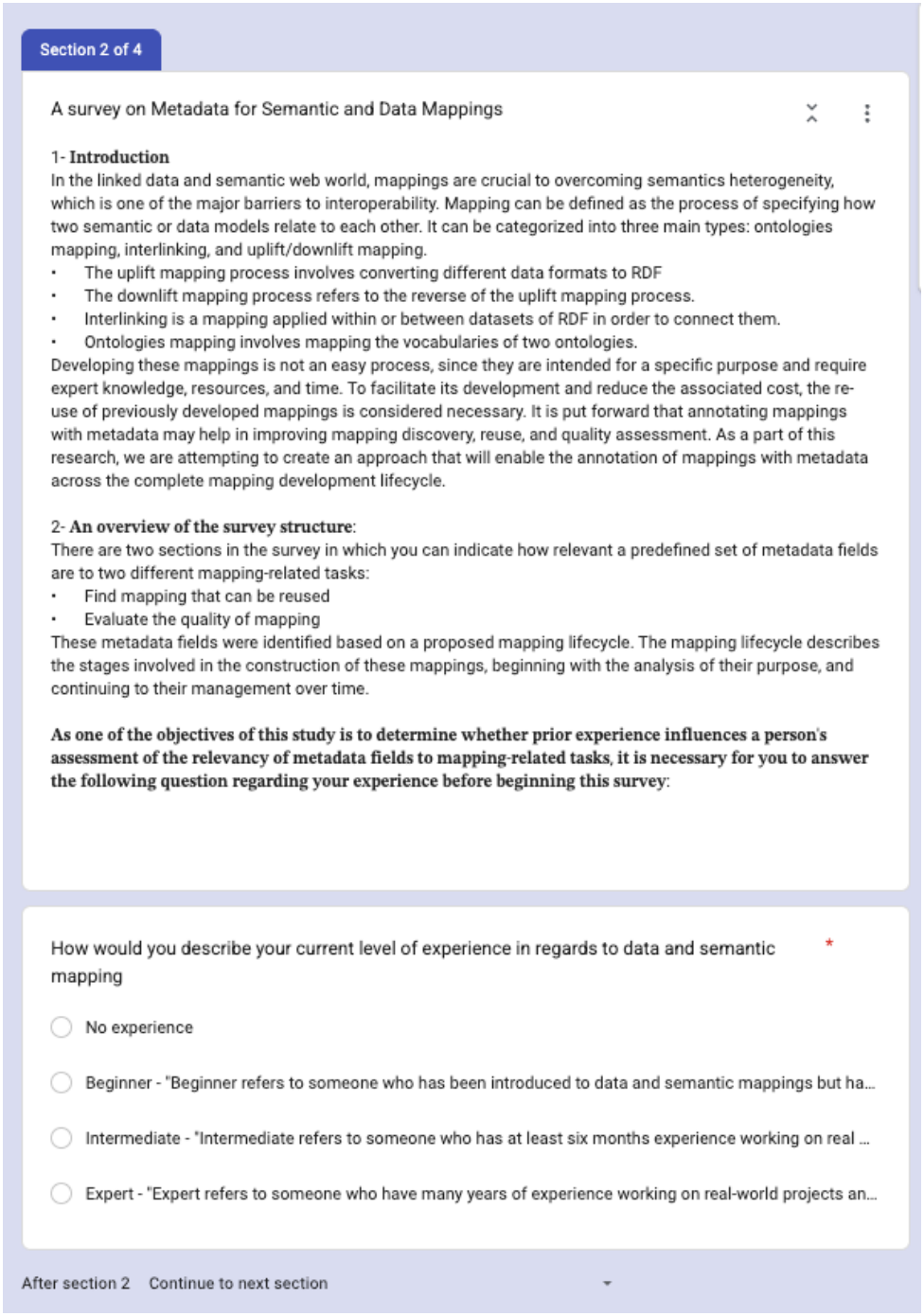
Please indicate your consent before proceeding *

- ○ I consent to participate in this study, and consent to the data processing necessary to enable my particip...
- ○ I don't consent

After section 1 Continue to next section

D.4: Survey instrument.

The survey instrument was administered using Google Forms and remains accessible at: https://forms.gle/22FU7kfqP8979AM37



Section 2 of 4

A survey on Metadata for Semantic and Data Mappings

1- **Introduction**
In the linked data and semantic web world, mappings are crucial to overcoming semantics heterogeneity, which is one of the major barriers to interoperability. Mapping can be defined as the process of specifying how two semantic or data models relate to each other. It can be categorized into three main types: ontologies mapping, interlinking, and uplift/downlift mapping.

- The uplift mapping process involves converting different data formats to RDF
- The downlift mapping process refers to the reverse of the uplift mapping process.
- Interlinking is a mapping applied within or between datasets of RDF in order to connect them.
- Ontologies mapping involves mapping the vocabularies of two ontologies.

Developing these mappings is not an easy process, since they are intended for a specific purpose and require expert knowledge, resources, and time. To facilitate its development and reduce the associated cost, the re-use of previously developed mappings is considered necessary. It is put forward that annotating mappings with metadata may help in improving mapping discovery, reuse, and quality assessment. As a part of this research, we are attempting to create an approach that will enable the annotation of mappings with metadata across the complete mapping development lifecycle.

2- **An overview of the survey structure**:
There are two sections in the survey in which you can indicate how relevant a predefined set of metadata fields are to two different mapping-related tasks:

- Find mapping that can be reused
- Evaluate the quality of mapping

These metadata fields were identified based on a proposed mapping lifecycle. The mapping lifecycle describes the stages involved in the construction of these mappings, beginning with the analysis of their purpose, and continuing to their management over time.

As one of the objectives of this study is to determine whether prior experience influences a person's assessment of the relevancy of metadata fields to mapping-related tasks, it is necessary for you to answer the following question regarding your experience before beginning this survey:

How would you describe your current level of experience in regards to data and semantic mapping *

- No experience
- Beginner - "Beginner refers to someone who has been introduced to data and semantic mappings but ha...
- Intermediate - "Intermediate refers to someone who has at least six months experience working on real ...
- Expert - "Expert refers to someone who have many years of experience working on real-world projects an...

After section 2 Continue to next section

Section 3 of 4

1- Find a mapping that can be re-used

The process of creating mappings involves a considerable amount of time, effort, and expertise. This can be mitigated by reusing previously developed mappings. However, looking for mappings before reusing them is an essential step. In the event that mappings are stored in one portal, reasoning over the metadata fields associated with the mappings is one means of finding appropriate mappings to reuse.

In this section, you will be asked based on your experience to rate the relevance of metadata fields to the mapping discovery process for the purpose of reuse. It is expected that relevant metadata fields will facilitate the discovery process when queried.

On scale of 1 to 5, Please rate the following metadata fields (in your opinion) based on how relevant they are to the mapping discovery process.

Description (optional)

Metadata about stakeholders (A stakeholder is anybody who has contributed to the development of a mapping, such as knowledge engineers and domain experts)

	1 Highly releva...	2	3	4	5 Irrelevant to ...
Stakeholder's ...	○	○	○	○	○
Stakeholder's ...	○	○	○	○	○
Role of stakeh...	○	○	○	○	○
Background on...	○	○	○	○	○
Stakeholder or...	○	○	○	○	○

Metadata related to the purpose of the mapping project

	1 Highly releva...	2	3	4	5 Irrelevant to ...
Mapping requir...	○	○	○	○	○
Type of mappi...	○	○	○	○	○
Mapping doma...	○	○	○	○	○
Domain assum...	○	○	○	○	○

Domain assum...	○	○	○	○	○
Any technical r...	○	○	○	○	○
Risks or issues...	○	○	○	○	○

Metadata to describe inputs that will be mapped (In the case of ontology mapping, inputs would be ontologies, datasets or entities for the interlinking type of mapping, and any source of data for uplift mapping.)

	1 Highly releva...	2	3	4	5 Irrelevant to ...
Input name in ...	○	○	○	○	○
Input URI (Unif...	○	○	○	○	○
Source of this i...	○	○	○	○	○
Input type (eg: ...	○	○	○	○	○
Input creator	○	○	○	○	○
Input format (e...	○	○	○	○	○

Metadata to describe the design process of the mapping. . During this process, a decision is made as to whether an existing mapping can be reused or if a new one should be developed.

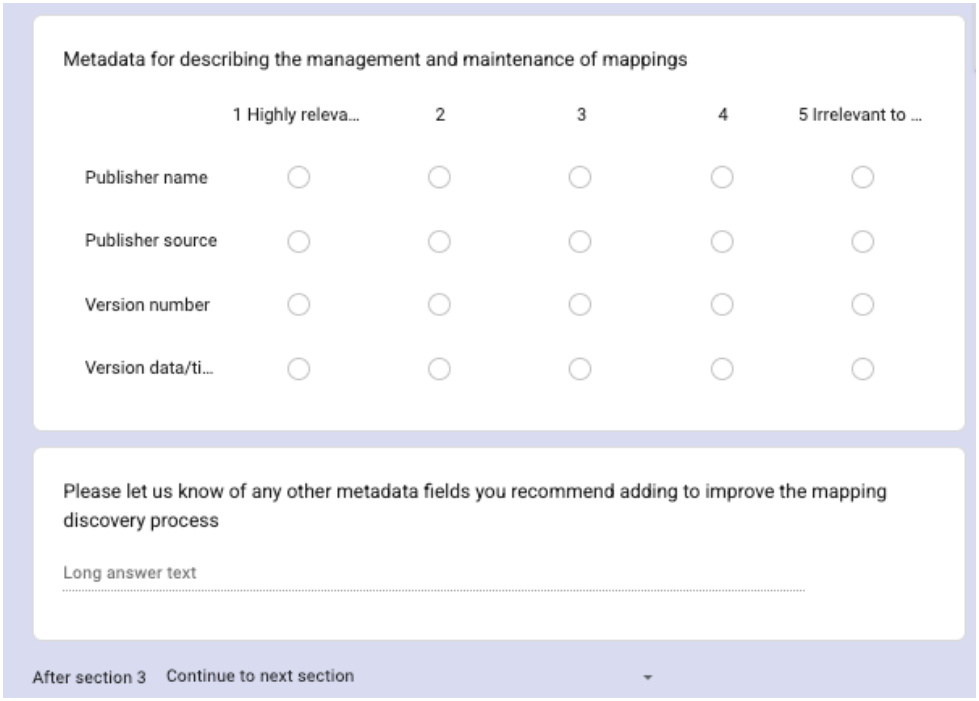
	1 Highly releva...	2	3	4	5 Irrelevant to ...
Final design de...	○	○	○	○	○
Design decisio...	○	○	○	○	○
Mapping qualit...	○	○	○	○	○

Metadata to describe the development process of a mapping

	1 Highly releva...	2	3	4	5 Irrelevant to ...
Mapping Huma...	○	○	○	○	○
Mapping URI (...	○	○	○	○	○
Development o...	○	○	○	○	○
Development o...	○	○	○	○	○
Tools used duri...	○	○	○	○	○
Mapping meth...	○	○	○	○	○
Mapping algori...	○	○	○	○	○
Mapping format	○	○	○	○	○

Metadata for describing the testing process of a mapping

	1 Highly releva...	2	3	4	5 Irrelevant to ...
Testing unique ...	○	○	○	○	○
Testing name	○	○	○	○	○
Testing Type (...	○	○	○	○	○
Testing Date/Ti...	○	○	○	○	○
Testing Result	○	○	○	○	○



Metadata for describing the management and maintenance of mappings

	1 Highly releva...	2	3	4	5 Irrelevant to ...
Publisher name	○	○	○	○	○
Publisher source	○	○	○	○	○
Version number	○	○	○	○	○
Version data/ti...	○	○	○	○	○

Please let us know of any other metadata fields you recommend adding to improve the mapping discovery process

Long answer text

After section 3 Continue to next section

Section 4 of 4

2- Evaluating the quality of mapping

Evaluating the quality of mapping is one of mapping related tasks. During this process, mapping quality should be assessed based on predefined quality metrics or criteria.

In this section, you will be asked based on your experience to rate the relevance of metadata fields to the quality assessment process. It is expected that relevant metadata fields will facilitate the quality assessment process when queried.

On scale of 1 to 5, Please rate the following metadata fields (in your opinion) based on how relevant they are to the mapping quality assessment process.

Description (optional)

Metadata about stakeholders (A stakeholder is anybody who has contributed to the development of a mapping, such as knowledge engineers and domain experts)

	1 Highly releva...	2	3	4	5 Irrelevant to ...
Stakeholder's ...	○	○	○	○	○
Stakeholder's ...	○	○	○	○	○
Role of stakeh...	○	○	○	○	○
Background on...	○	○	○	○	○
Stakeholder or...	○	○	○	○	○

Metadata related to the purpose of this mapping project

	1 Highly releva...	2	3	4	5 Irrelevant to ...
Mapping requir...	○	○	○	○	○
Type of mappi...	○	○	○	○	○
Mapping doma...	○	○	○	○	○
Domain assum...	○	○	○	○	○
Any technical r...	○	○	○	○	○
Risks or issues...	○	○	○	○	○

Metadata to describe inputs that will be mapped (In the case of ontology mapping, inputs would be ontologies, datasets or entities for the interlinking type of mapping, and any source of data for uplift mapping.)

	1 Highly releva...	2	3	4	5 Irrelevant to ...
Input name in ...	○	○	○	○	○
Input URI (Unif...	○	○	○	○	○
Source of this i...	○	○	○	○	○
Input type (eg: ...	○	○	○	○	○
Input creator	○	○	○	○	○
Input format (e...	○	○	○	○	○

Metadata to describe the design process of the mapping. . During this process, a decision is made as to whether an existing mapping can be reused or if a new one should be developed.

	1 Highly releva...	2	3	4	5 Irrelevant to ...
Final design de...	○	○	○	○	○
Design decisio...	○	○	○	○	○
Mapping qualit...	○	○	○	○	○

Metadata to describe the development process of a mapping

	1 Highly releva...	2	3	4	5 Irrelevant to ...
Mapping Huma...	○	○	○	○	○
Mapping URI (...	○	○	○	○	○
Development o...	○	○	○	○	○
Development o...	○	○	○	○	○
Tools used duri...	○	○	○	○	○
Mapping meth...	○	○	○	○	○
Mapping algori...	○	○	○	○	○
Mapping format	○	○	○	○	○

Metadata for describing the testing process of a mapping

	1 Highly releva...	2	3	4	5 Irrelevant to ...
Testing Unique...	○	○	○	○	○
Testing name	○	○	○	○	○
Testing Type (...	○	○	○	○	○
Testing Date/Ti...	○	○	○	○	○
Testing Result	○	○	○	○	○

Metadata for describing the management and maintenance of mappings

	1 Highly releva...	2	3	4	5 Irrelevant to ...
Publisher name	○	○	○	○	○
Publisher source	○	○	○	○	○
Version number	○	○	○	○	○
Version data/ti...	○	○	○	○	○

Finally, if you are aware of any other metadata fields that would improve assessing the quality of mappings, please let us know about them.

Long answer text

D.5: Round 2 Results tables

Table D.1 Ratings results of the metadata fields relevance to mapping discovery task- Follow-up Round Intermediates

Mapping discovery task - Follow-up Round Intermediates (n=3)						
	Proposed fields	**1 Highly relevant**	**2**	**3**	**4**	**5 Irrelevant**
Metadata fields describing stakeholders	Name	1	1	1	0	0
	URI	0	3	0	0	0
	Role	1	2	0	0	0
	Background	2	1	0	0	0
	Organization	0	2	0	1	0
Metadata fields describing the mapping project's purpose	Requirement	2	1	0	0	0
	Type of mapping	1	2	0	0	0
	Domain	2	1	0	0	0
	Assumptions	1	1	1	0	0
	Technical Requirement	1	1	0	1	0
	Risks/issues	1	2	0	0	0
Metadata fields describing the input to be mapped	Name	1	1	1	0	0
	URI	0	3	0	0	0
	Source	1	2	0	0	0
	Type	0	3	0	0	0
	Creator	1	2	0	0	0
	Format	1	2	0	0	0
Metadata fields describing the mapping design process	Design decisions	2	1	0	0	0
	Justifications	1	2	0	0	0
	Quality Metrics	3	0	0	0	0
Metadata fields describing the mapping development process	Mapping Name	3	0	0	0	0
	Mapping URI	2	1	0	0	0
	Start date and time	3	0	0	0	0
	End date and time	3	0	0	0	0
	Used tools	0	2	1	0	0
	Mapping method	3	0	0	0	0
	Mapping algorithm	3	0	0	0	0
	Mapping format	1	2	0	0	0
Metadata fields describing the testing process of a mapping	Testing Identifier	1	2	0	0	0
	Testing Name	0	2	1	0	0
	Testing type	3	0	0	0	0
	Testing data and time	2	1	0	0	0
	Testing result	2	1	0	0	0
Metadata fields describing the management and maintenance of mappings	Publisher name	1	2	0	0	0
	Publisher source	3	0	0	0	0
	Version number	3	0	0	0	0
	Version date/time	2	1	0	0	0

Table D.2 Ratings results of the metadata fields relevance to mapping discovery task - Follow-up Round Beginners

Mapping discovery task - Follow-up Round Beginners (n=46)						
	Proposed fields	**1 Highly relevant**	**2**	**3**	**4**	**5 Irrelevant**
Metadata fields describing stakeholders	Name	11	15	9	8	2
	URI	23	13	5	4	0
	Role	9	16	11	7	2
	BacTwo further improvements were directly traceable to specific MMV fields.round	5	7	14	13	6
	Organization	10	14	17	1	3
Metadata fields describing the mapping project's purpose	Requirement	27	16	3	0	0
	Type of mapping	26	15	5	0	0
	Domain	22	17	6	0	1
	Assumptions	16	17	10	3	0
	Technical Requirement	10	16	13	6	1
	Risks/issues	9	18	13	5	1
Metadata fields describing the input to be mapped	Name	13	17	10	6	0
	URI	24	16	4	2	0
	Source	12	20	11	3	0
	Type	18	17	8	3	0
	Creator	7	11	15	9	3
	Format	15	14	7	8	1
Metadata fields describing the mapping design process	Design decisions	21	20	3	0	1
	Justifications	15	15	13	2	0
	Quality Metrics	20	15	7	3	0
Metadata fields describing the mapping development process	Mapping Name	11	17	12	6	0
	Mapping URI	24	12	8	2	0
	Start date and time	7	12	16	7	4
	End date and time	6	15	15	5	4
	Used tools	8	18	13	6	1
	Mapping method	14	17	10	5	0
	Mapping algorithm	20	17	5	4	0
	Mapping format	18	18	8	2	0
Metadata fields describing the testing process of a mapping	Testing Identifier	20	13	6	6	1
	Testing Name	8	18	11	8	0
	Testing type	15	17	7	7	0
	Testing data and time	8	14	11	10	3
	Testing result	23	14	4	5	0
Metadata fields describing the management and maintenance of mappings	Publisher name	12	14	13	6	1
	Publisher source	18	13	7	8	0
	Version number	17	16	8	4	1
	Version date/time	16	13	8	8	1

Table D.3 Ratings results of the metadata fields relevance to quality assessment task- Follow-up Round Intermediates

Mapping quality assessment task - Follow-up Round Intermediates (n=3)						
	Proposed fields	**1 Highly relevant**	**2**	**3**	**4**	**5 Irrelevant**
Metadata fields describing stakeholders	Name	0	3	0	0	0
	URI	1	2	0	0	0
	Role	3	0	0	0	0
	Background	1	2	0	0	0
	Organization	2	0	1	0	0
Metadata fields describing the mapping project's purpose	Requirement	3	0	0	0	0
	Type of mapping	2	1	0	0	0
	Domain	2	1	0	0	0
	Assumptions	2	1	0	0	0
	Technical Requirement	2	1	0	0	0
	Risks/issues	1	1	1	0	0
Metadata fields describing the input to be mapped	Name	0	3	0	0	0
	URI	2	1	0	0	0
	Source	3	0	0	0	0
	Type	0	2	0	0	0
	Creator	2	0	1	0	0
	Format	1	2	0	0	0
Metadata fields describing the mapping design process	Design decisions	1	2	0	0	0
	Justifications	2	1	0	0	0
	Quality Metrics	2	1	0	0	0
Metadata fields describing the mapping development process	Mapping Name	1	2	0	0	0
	Mapping URI	1	2	0	0	0
	Start date and time	2	1	0	0	0
	End date and time	1	2	0	0	0
	Used tools	2	1	0	0	0
	Mapping method	3	0	0	0	0
	Mapping algorithm	2	1	0	0	0
	Mapping format	1	2	0	0	0
Metadata fields describing the testing process of a mapping	Testing Identifier	1	2	0	0	0
	Testing Name	2	0	1	0	0
	Testing type	2	1	0	0	0
	Testing data and time	2	0	1	0	0
	Testing result	3	0	0	0	0
Metadata fields describing the management and maintenance of mappings	Publisher name	2	1	0	0	0
	Publisher source	3	0	0	0	0
	Version number	1	2	0	0	0
	Version date/time	2	1	0	0	0

Table D.4 Ratings results of the metadata fields relevance to quality assessment task- Follow-up Round Beginners

Mapping quality assessment task - Follow-up Round Beginners (n=46)						
	Proposed fields	**1 Highly relevant**	**2**	**3**	**4**	**5 Irrelevant**
Metadata fields describing stakeholders	Name	13	14	8	8	3
	URI	19	14	10	3	0
	Role	11	19	9	3	4
	Background	6	13	11	11	5
	Organization	10	11	15	8	2
Metadata fields describing the mapping project's purpose	Requirement	24	13	7	1	0
	Type of mapping	23	15	4	3	0
	Domain	18	15	8	2	2
	Assumptions	21	14	6	3	1
	Technical Requirement	11	20	9	4	1
	Risks/issues	13	16	9	5	2
Metadata fields describing the input to be mapped	Name	15	12	12	3	4
	URI	23	14	7	2	0
	Source	19	14	10	3	0
	Type	24	12	6	4	0
	Creator	10	8	19	7	2
	Format	14	16	11	5	0
Metadata fields describing the mapping design process	Design decisions	22	16	6	2	0
	Justifications	13	20	8	3	2
	Quality Metrics	23	16	4	3	0
Metadata fields describing the mapping development process	Mapping Name	15	13	6	9	2
	Mapping URI	22	14	3	6	0
	Start date and time	6	14	11	9	5
	End date and time	6	12	13	9	5
	Used tools	10	14	9	8	4
	Mapping method	15	14	10	5	1
	Mapping algorithm	22	14	5	3	1
	Mapping format	20	12	11	1	1
Metadata fields describing the testing process of a mapping	Testing Identifier	23	11	7	5	0
	Testing Name	9	13	14	8	1
	Testing type	21	16	2	5	1
	Testing data and time	10	17	7	8	4
	Testing result	27	10	5	4	0
Metadata fields describing the management and maintenance of mappings	Publisher name	7	22	6	9	2
	Publisher source	16	16	5	8	1
	Version number	19	12	9	4	2
	Version date/time	17	13	5	6	4

Appendix E: Study 2 Materials

E.2: Invitation Email

Dear participant,

I hope you are doing well,

I am a Ph.D. student at Trinity College Dublin, Ireland, conducting research on metadata for semantic/data mappings. As part of this research, I am organizing an experiment to gather insights into mapping annotation processes.

Background:
In the linked data and semantic web communities, mappings play a critical role in addressing semantics heterogeneity, which is a significant barrier to interoperability. Mapping annotation involves enriching mappings with metadata to enhance their discoverability and reusability.

Purpose of the Experiment:
I am inviting you to participate in this experiment, where you will be asked to annotate mappings with a set of metadata fields. Your participation will contribute to refining tools for mapping discovery and reusability, ultimately benefiting the semantic web and linked data communities.

Duration and Participation:
Participation in this experiment should take no longer than thirty minutes of your time. Your input will be invaluable in improving mapping annotation processes.

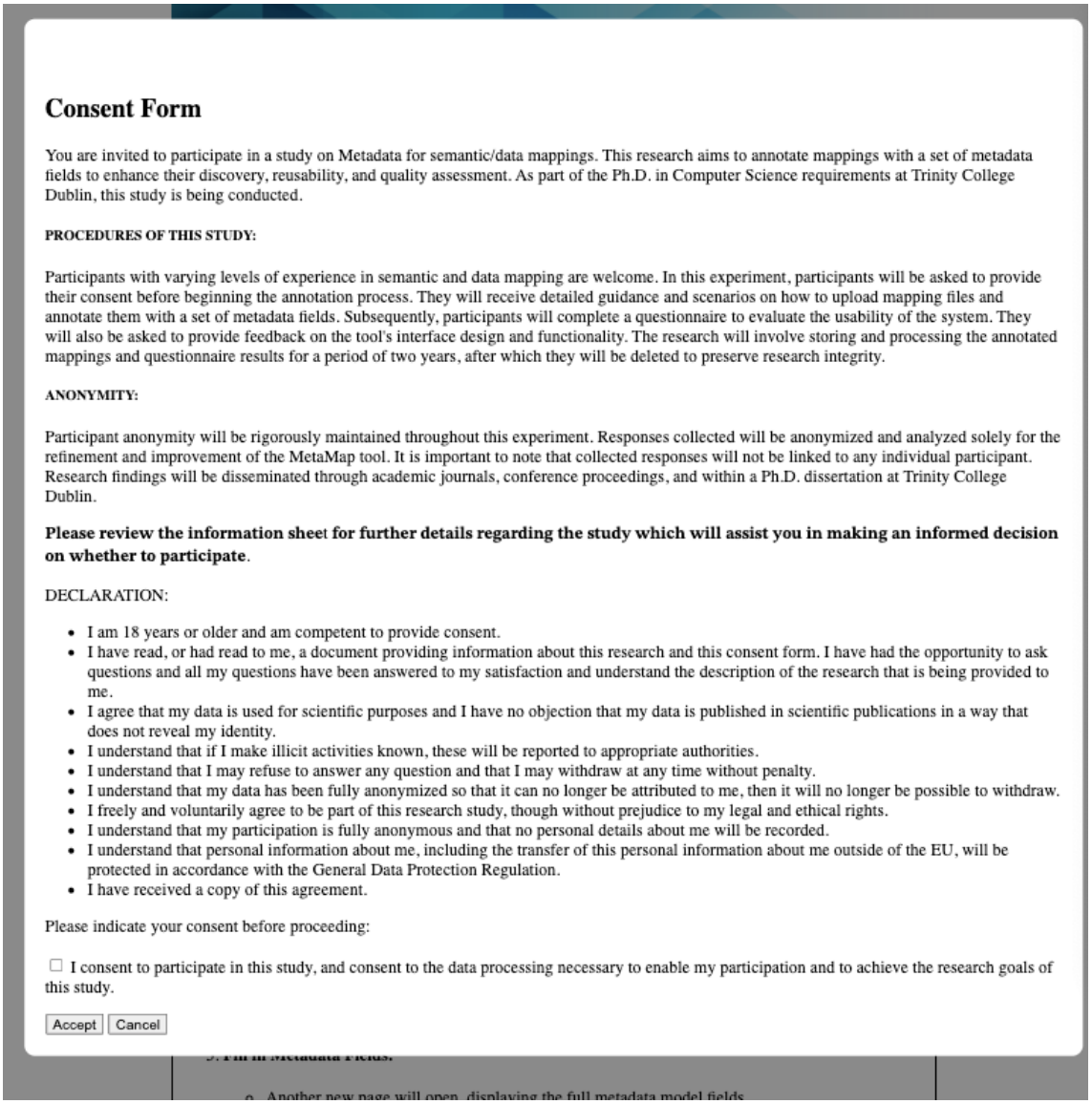
Confidentiality and Ethics Approval:
Your participation is entirely voluntary, and all responses will be anonymized. None of the responses will be linked to identifying information. This experiment has been approved by the School of Computer Science and Statistics ethics committee at Trinity College Dublin.

Experiment Link: To participate, please click on the following link: https://metasemap-1-sarah-alzahrani.replit.app/

Contact Information:
If you have any questions or encounter any difficulties during the experiment, please feel free to contact me at salzahra@tcd.ie.
Your participation is highly appreciated. Thank you for considering contributing to this research endeavor.

Many Thanks,
Sarah Alzahrani
Trinity College Dublin

E.2: Consent form



Consent Form

You are invited to participate in a study on Metadata for semantic/data mappings. This research aims to annotate mappings with a set of metadata fields to enhance their discovery, reusability, and quality assessment. As part of the Ph.D. in Computer Science requirements at Trinity College Dublin, this study is being conducted.

PROCEDURES OF THIS STUDY:

Participants with varying levels of experience in semantic and data mapping are welcome. In this experiment, participants will be asked to provide their consent before beginning the annotation process. They will receive detailed guidance and scenarios on how to upload mapping files and annotate them with a set of metadata fields. Subsequently, participants will complete a questionnaire to evaluate the usability of the system. They will also be asked to provide feedback on the tool's interface design and functionality. The research will involve storing and processing the annotated mappings and questionnaire results for a period of two years, after which they will be deleted to preserve research integrity.

ANONYMITY:

Participant anonymity will be rigorously maintained throughout this experiment. Responses collected will be anonymized and analyzed solely for the refinement and improvement of the MetaMap tool. It is important to note that collected responses will not be linked to any individual participant. Research findings will be disseminated through academic journals, conference proceedings, and within a Ph.D. dissertation at Trinity College Dublin.

Please review the information sheet for further details regarding the study which will assist you in making an informed decision on whether to participate.

DECLARATION:

- I am 18 years or older and am competent to provide consent.
- I have read, or had read to me, a document providing information about this research and this consent form. I have had the opportunity to ask questions and all my questions have been answered to my satisfaction and understand the description of the research that is being provided to me.
- I agree that my data is used for scientific purposes and I have no objection that my data is published in scientific publications in a way that does not reveal my identity.
- I understand that if I make illicit activities known, these will be reported to appropriate authorities.
- I understand that I may refuse to answer any question and that I may withdraw at any time without penalty.
- I understand that my data has been fully anonymized so that it can no longer be attributed to me, then it will no longer be possible to withdraw.
- I freely and voluntarily agree to be part of this research study, though without prejudice to my legal and ethical rights.
- I understand that my participation is fully anonymous and that no personal details about me will be recorded.
- I understand that personal information about me, including the transfer of this personal information about me outside of the EU, will be protected in accordance with the General Data Protection Regulation.
- I have received a copy of this agreement.

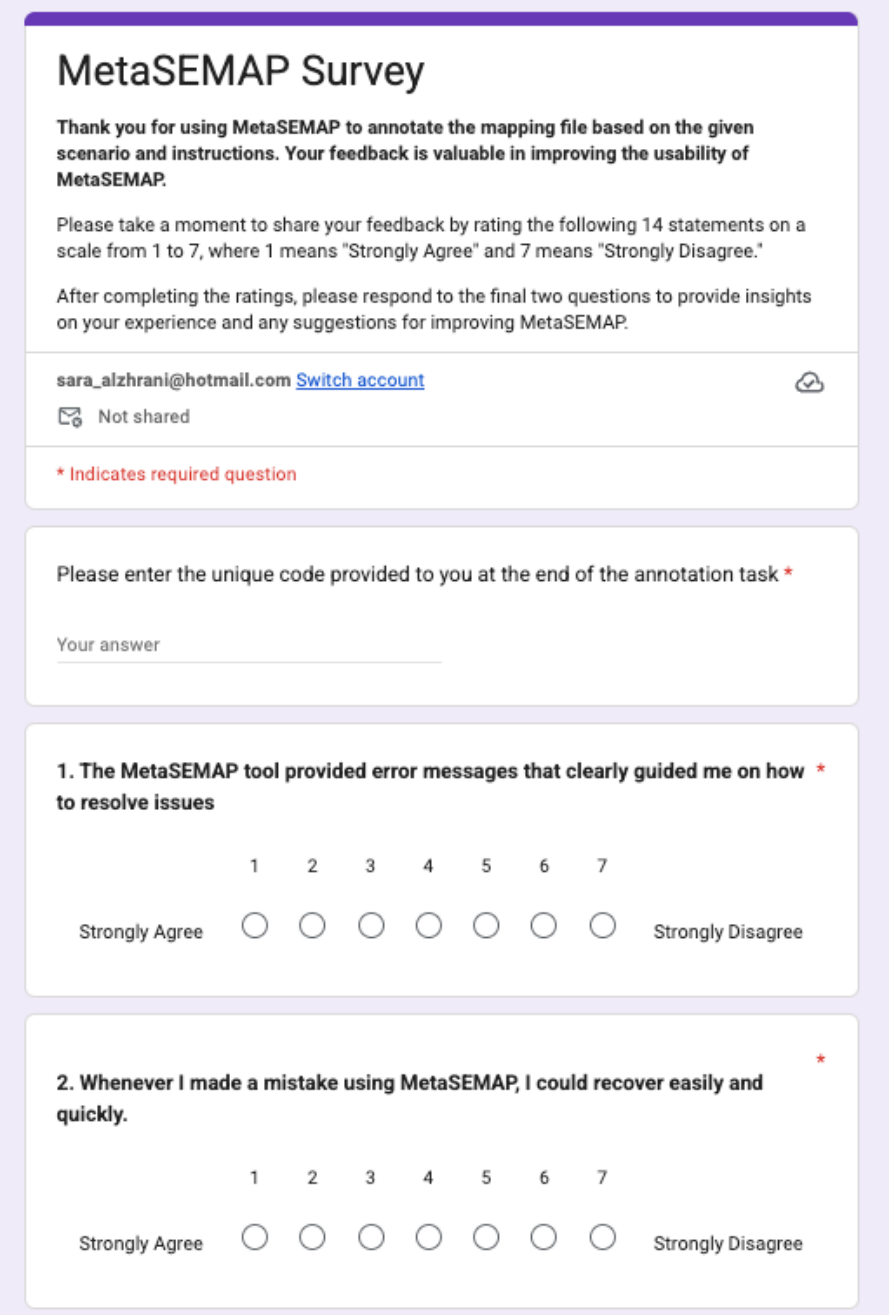
Please indicate your consent before proceeding:

☐ I consent to participate in this study, and consent to the data processing necessary to enable my participation and to achieve the research goals of this study.

Accept Cancel

E.3: Questionnaire

The survey instrument was administered using Google Forms and remains accessible at: https://forms.gle/rYmyQm2AbTZdYGqdA



MetaSEMAP Survey

Thank you for using MetaSEMAP to annotate the mapping file based on the given scenario and instructions. Your feedback is valuable in improving the usability of MetaSEMAP.

Please take a moment to share your feedback by rating the following 14 statements on a scale from 1 to 7, where 1 means "Strongly Agree" and 7 means "Strongly Disagree."

After completing the ratings, please respond to the final two questions to provide insights on your experience and any suggestions for improving MetaSEMAP.

sara_alzhrani@hotmail.com Switch account

Not shared

* Indicates required question

Please enter the unique code provided to you at the end of the annotation task *

Your answer

1. The MetaSEMAP tool provided error messages that clearly guided me on how to resolve issues *

	1	2	3	4	5	6	7	
Strongly Agree	○	○	○	○	○	○	○	Strongly Disagree

2. Whenever I made a mistake using MetaSEMAP, I could recover easily and quickly. *

	1	2	3	4	5	6	7	
Strongly Agree	○	○	○	○	○	○	○	Strongly Disagree

4. MetaSEMAP provides all the functions I need for annotating mapping files. *

	1	2	3	4	5	6	7	
Strongly Agree	○	○	○	○	○	○	○	Strongly Disagree

5. MetaSEMAP is easy to use for annotating mapping files. *

	1	2	3	4	5	6	7	
Strongly Agree	○	○	○	○	○	○	○	Strongly Disagree

6. It is easy to learn how to use MetaSEMAP for annotating mapping files. *

	1	2	3	4	5	6	7	
Strongly Agree	○	○	○	○	○	○	○	Strongly Disagree

7. I can imagine that most people would learn to use MetaSEMAP for annotating mapping files very quickly. *

	1	2	3	4	5	6	7	
Strongly Agree	○	○	○	○	○	○	○	Strongly Disagree

8. I find MetaSEMAP unnecessarily complex for annotating mapping files. *

	1	2	3	4	5	6	7	
Strongly Agree	○	○	○	○	○	○	○	Strongly Disagree

9. I needed to learn a lot of things before I could effectively use MetaSEMAP for annotating mapping files. *

	1	2	3	4	5	6	7	
Strongly Agree	◯	◯	◯	◯	◯	◯	◯	Strongly Disagree

10. The various functions in MetaSEMAP are well-integrated for annotating mapping files. *

	1	2	3	4	5	6	7	
Strongly Agree	◯	◯	◯	◯	◯	◯	◯	Strongly Disagree

11. I thought there was too much inconsistency in MetaSEMAP for annotating mapping files. *

	1	2	3	4	5	6	7	
Strongly Agree	◯	◯	◯	◯	◯	◯	◯	Strongly Disagree

12. I felt very confident using MetaSEMAP for annotating mapping files. *

	1	2	3	4	5	6	7	
Strongly Agree	◯	◯	◯	◯	◯	◯	◯	Strongly Disagree

13. I found MetaSEMAP challenging to use for annotating mapping files. *

	1	2	3	4	5	6	7	
Strongly Agree	○	○	○	○	○	○	○	Strongly Disagree

14. Overall, I am satisfied with my experience using MetaSEMAP for annotating mapping files. *

	1	2	3	4	5	6	7	
Strongly Agree	○	○	○	○	○	○	○	Strongly Disagree

Which metadata representation in MetaSEMAP did you find easiest to read and understand when annotating mapping files? *

- ○ Named Graph
- ○ RDF-star

Do you have any additional comments or suggestions on how we can improve the design of MetaSEMAP for annotating mapping files? Please share your thoughts below. *

Your answer

Submit Clear form

Appendix F: Study 3 Materials

F.1: Ethics application approval letter:

25/07/2025

Title: Evaluation of a Lifecycle-Based Metadata Model

REAMs No: 4795

Dear Sarah Alzahrani,

Your application has been reviewed by the relevant Research Ethics Committee (REC) and we are pleased to inform you that the above project has been approved.

Please note that any documents submitted for GDPR purposes in association with your ethics application are approved by the REC from an ethical perspective only (GDPR compliance is assumed, as per your uploaded letter of approval from the DPO if applicable).

It is the responsibility of the researcher/research team to ensure that, if applicable, all aspects of the study are executed in compliance with the General Data Protection Regulation (GDPR), the Health Research Regulations, Data Protection Act 2018, particularly regarding the storage and destruction of data arising from the research, and adhere to any relevant Health & Safety regulations.

Please note the reporting requirements outlined, in particular the need for:

- An immediate report of any serious or unexpected adverse event should be sent immediately to your REC via email. The Procedures for Reporting Adverse Events are outlined in the application. The Adverse Event form are available here.
- Any other unforeseen events should also be reported using this form and sent to your REC via email. The Adverse Event form are available here.
- An end of project report is due one year from the end of project date. The form is available here. This should be sent to your REC via email.
- If the data collection period is longer than 1 year, an annual report should be submitted. The form is available here. This should be sent to your REC via email.

You, and your research team (where applicable), are responsible for conducting the research strictly in the manner outlined in your application and supporting documentation, and in compliance with it. Proposed changes or amendments to the approved study must be submitted to the REC for approval in advance of any changes or amendments being implemented. In the event that the changes or amendments are significant, the REC may require a new submission. This ethical approval expires with the end dates of the project as stated in the application.

If you have any queries regarding this decision, please contact the Chair of your REC.

We wish you all the very best with your research project.

Kind Regards,

The REC

NB. If your study is to be conducted outside TCD, e.g. in a hospital site, please note that Trinity College REC approval does not of itself entitle you to conduct research in an external site. In this regard, the onus is upon you - the researcher, to ensure that all permissions have been obtained, and that any other requirements of the study site have been satisfied

NB. If your study involves a trial of medicinal products or medical devices, apps on humans you must provide a copy of your Clinical Trial Sponsorship from TCD - indicating that you have the appropriate trial insurance and sponsor oversigh

F.2: Invitation Email:

Dear XXXX,

I hope you are well,

I am conducting a research study as part of my PhD at Trinity College Dublin, under the supervision of Professor Declan O'Sullivan, examining a lifecycle-based metadata ontology designed to improve the documentation, transparency, and reuse of declarative mappings (e.g., RML mappings, ontology alignments, and interlinking rules).

I am contacting you in your capacity as an expert in knowledge graphs, semantic web technologies, and mapping engineering, as your evaluation would provide valuable insight into the validity, completeness, and practical applicability of the proposed metadata model.

The study involves reviewing a short ontology specification and completing an online evaluation survey. The survey focuses on the clarity, structure, completeness, and perceived usefulness of the metadata model. It takes approximately 15–20 minutes to complete and can be done at your convenience. Your feedback would be greatly appreciated in refining the ontology and strengthening the evaluation of this work.

If you are willing to participate, you can access the survey here:

https://forms.gle/7Yw7pRRSxbWNSV639

The deadline for responses is Wednesday, 25th February, but if you need a little extra time, please just email me to request so, as your expert opinion is really important for my study.

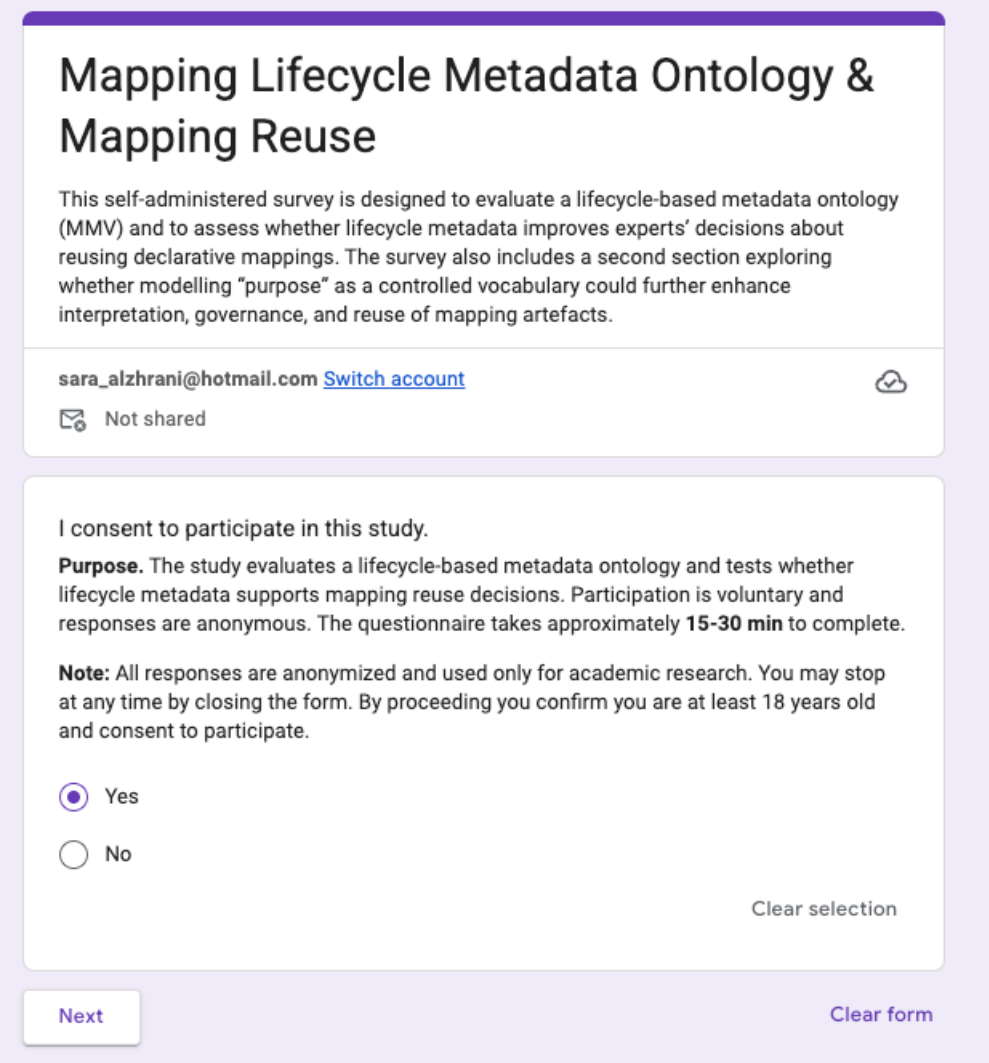
Participation is voluntary, and the survey is anonymous. Responses will be used solely for academic research purposes.

Please feel free to contact me if you have any questions about the study.

Many thanks in advance for considering this request.

Best regards,
Sara Alzahrani
PhD Researcher
Trinity College Dublin

F.3: Consent Form:



Mapping Lifecycle Metadata Ontology & Mapping Reuse

This self-administered survey is designed to evaluate a lifecycle-based metadata ontology (MMV) and to assess whether lifecycle metadata improves experts' decisions about reusing declarative mappings. The survey also includes a second section exploring whether modelling "purpose" as a controlled vocabulary could further enhance interpretation, governance, and reuse of mapping artefacts.

sara_alzhrani@hotmail.com Switch account

Not shared

I consent to participate in this study.

Purpose. The study evaluates a lifecycle-based metadata ontology and tests whether lifecycle metadata supports mapping reuse decisions. Participation is voluntary and responses are anonymous. The questionnaire takes approximately **15-30 min** to complete.

Note: All responses are anonymized and used only for academic research. You may stop at any time by closing the form. By proceeding you confirm you are at least 18 years old and consent to participate.

- (●) Yes
- ( ) No

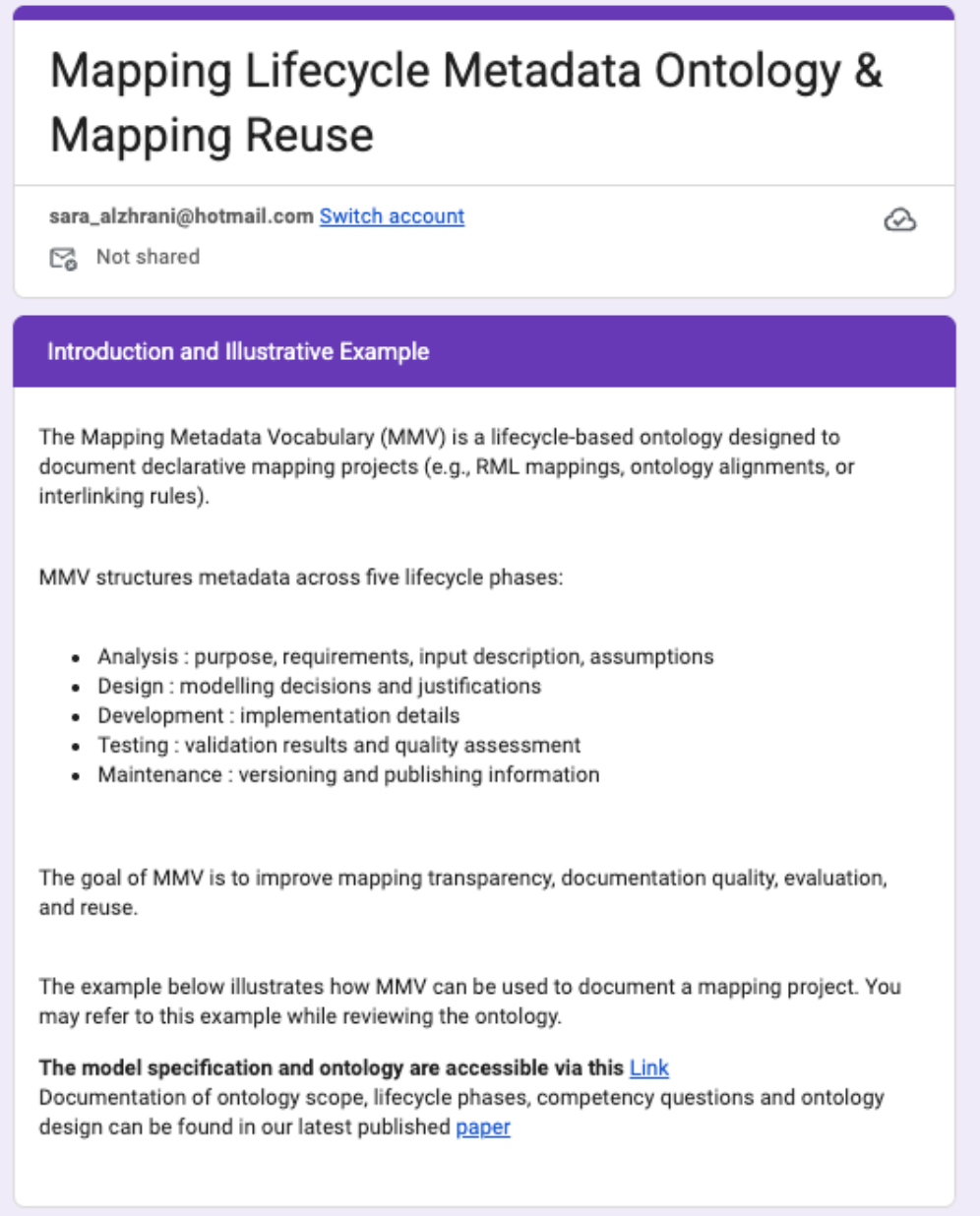
Clear selection

Next

Clear form

F.4: expert evaluation questionnaire

The survey instrument was administered using Google Forms and remains accessible at: https://forms.gle/KRup2j6CYLzP5mPe9



Mapping Lifecycle Metadata Ontology & Mapping Reuse

sara_alzhrani@hotmail.com Switch account

Not shared

Introduction and Illustrative Example

The Mapping Metadata Vocabulary (MMV) is a lifecycle-based ontology designed to document declarative mapping projects (e.g., RML mappings, ontology alignments, or interlinking rules).

MMV structures metadata across five lifecycle phases:

- Analysis : purpose, requirements, input description, assumptions
- Design : modelling decisions and justifications
- Development : implementation details
- Testing : validation results and quality assessment
- Maintenance : versioning and publishing information

The goal of MMV is to improve mapping transparency, documentation quality, evaluation, and reuse.

The example below illustrates how MMV can be used to document a mapping project. You may refer to this example while reviewing the ontology.

The model specification and ontology are accessible via this Link
Documentation of ontology scope, lifecycle phases, competency questions and ontology design can be found in our latest published paper

Illustrative MMV example (Turtle representation).

The example below shows a mapping project documented using MMV. It demonstrates how a MappingProject links to lifecycle phases (Analysis, Design, Testing, Maintenance). Each phase records specific metadata (e.g., purpose, design decisions, testing results). Stakeholders and quality metrics are explicitly represented. In this example:

- The project (ex:Project1) is connected to its lifecycle phases.

- The Analysis phase records the mapping purpose and input dataset.

- The Design phase records modelling decisions and planned quality metrics.

- The Testing phase captures validation results.

- The Maintenance phase records publishing information.

You do not need to analyse the syntax in detail. The example is provided to illustrate how MMV structures mapping metadata.

```
@prefix mmv: <http://w3id.org/mmv/0.1#> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix dqv: <http://www.w3.org/ns/dqv#> .
@prefix ex: <http://example.org/> .

ex:Project1 a mmv:MappingProject ;
    mmv:hasAnalysis ex:Analysis1 ;
    mmv:hasDesign ex:Design1 ;
    mmv:hasTesting ex:Test1 ;
    mmv:hasMaintenance ex:Maintenance1 ;
    mmv:mmvhasStakeholder ex:Sarah .

ex:Analysis1 a mmv:AnalysisActivity ;
    mmv:hasPurpose "Transform CSV dataset into RDF for publication" ;
    mmv:hasMappingDomain "Geospatial" ;
    mmv:hasRequirement "Comply with GeoSPARQL" ;
    mmv:hasInputName "ElectoralDivisions2022.csv" ;
    mmv:hasAssumption "Coordinates are provided in WGS84" .

ex:Design1 a mmv:DesignActivity ;
    mmv:hasDesignDecision "Reuse GeoSPARQL vocabulary for geometry representation" ;
    mmv:hasJustification "Ensures interoperability with geospatial Linked Data tools" ;
    mmv:hasPlannedQualityMetric ex:CompletenessMetric .

ex:CompletenessMetric a dqv:Metric .

ex:Test1 a mmv:TestingActivity ;
    mmv:hasTestingType "SHACL validation" ;
    mmv:hasTestingResult "All constraints passed" .

ex:Maintenance1 a mmv:MaintenanceActivity ;
    mmv:hasPublisherName "Trinity College Dublin" .

ex:Sarah a prov:Agent ;
    mmv:hasRole "Mapping Developer" .
```

Based on the description, example, and documentation provided, please indicate your level of agreement with the following statements.

1- The ontology captures the essential concepts and relationships required to document mapping projects across the lifecycle.

	1	2	3	4	5	
Strongly Agree	○	○	○	○	○	Strongly Disagree

2- The overall structure of the ontology is coherent and easy to understand.

	1	2	3	4	5	
Strongly Agree	○	○	○	○	○	Strongly Disagree

3- The ontology classes and properties are clearly defined and non-ambiguous.

	1	2	3	4	5	
Strongly Agree	○	○	○	○	○	Strongly Disagree

4- The ontology follows recognised ontology design principles (e.g., clarity, coherence, reuse).

	1	2	3	4	5	
Strongly Agree	○	○	○	○	○	Strongly Disagree

5- The ontology appropriately reuses existing and well-established vocabularies.

	1	2	3	4	5	
Strongly Agree	○	○	○	○	○	Strongly Disagree

6- The ontology would support informed and confident decisions about mapping reuse.

	1	2	3	4	5	
Strongly Agree	○	○	○	○	○	Strongly Disagree

7- Do you suggest any modifications to the concepts, properties or modelling decisions ? Please justify.

Your answer

8- Do you suggest improvements to the documentation (examples, explanations, constraints)? Please justify your response.

Your answer

Back Next Clear form

Purpose Vocabulary

We are exploring whether "purpose" should be modelled as a first-class metadata element supported by a controlled vocabulary to improve reuse, interpretation, and governance of mapping artefacts. The following questions explore whether modelling "purpose" explicitly could strengthen mapping interpretation and reuse.

1- Do you think the purpose of a mapping artefact should be explicitly captured using a controlled vocabulary?

- ◯ Yes, strongly agree
- ◯ Yes, agree
- ◯ Neutral / unsure
- ◯ Disagree
- ◯ Strongly disagree

2- Which of the following purpose categories do you consider useful for describing mapping artefacts *(Select all that apply)*

- ☐ 1- Project Goal (e.g., data integration, enrichment)
- ☐ 2- Application Context or Domain (e.g., research, operational system, decision support)
- ☐ 3- Intended Use (e.g., analysis, publication, data exchange)
- ☐ 4- Target Audience (e.g., domain experts, developers, public users)
- ☐ 5- Reuse Expectations (e.g., exploratory reuse, production reuse, one-off use)

3- Rank the importance of the following purpose categories (1 = Most Important, 5 = Least Important)

	1	2	3	4	5
Project Goal	◯	◯	◯	◯	◯
Application Context or Domain	◯	◯	◯	◯	◯
Intended Use	◯	◯	◯	◯	◯
Target Audience	◯	◯	◯	◯	◯
Reuse Expectations	◯	◯	◯	◯	◯

4- Do you have suggestions for additional purpose categories, or comments on how purpose metadata could support reuse and interpretation of mappings?

Your answer

5- Do you think a controlled vocabulary for describing the purpose of an artefact should be applicable beyond mapping projects, to Semantic Web projects in general (e.g., datasets, ontologies, workflows)?

- ◯ Yes, strongly agree
- ◯ Yes, agree
- ◯ Neutral / unsure
- ◯ Disagree
- ◯ Strongly disagree

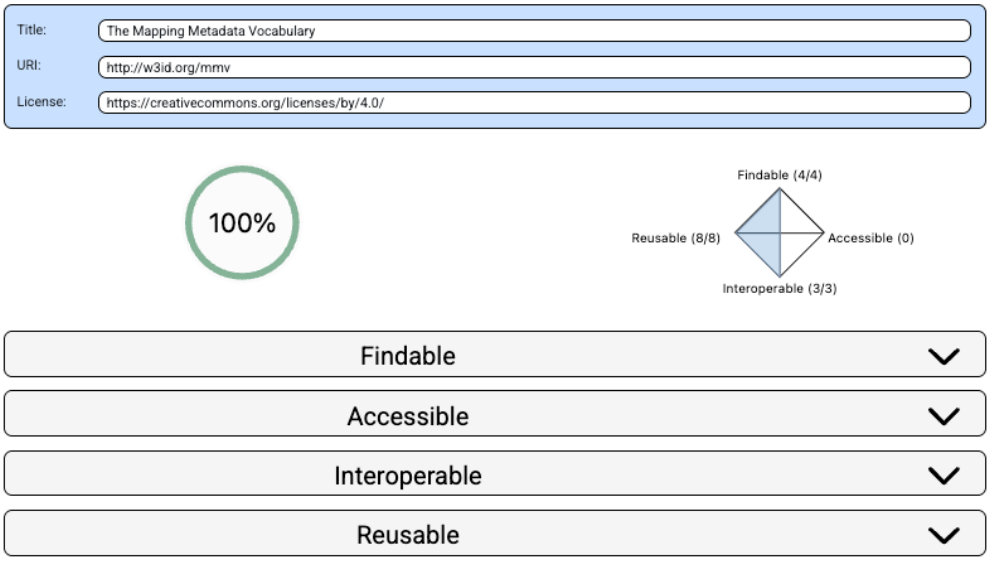
6- if yes, please describe how a general purpose-oriented vocabulary could be used in practice (e.g., to support reuse, interpretation, governance, or decision-making).

Your answer

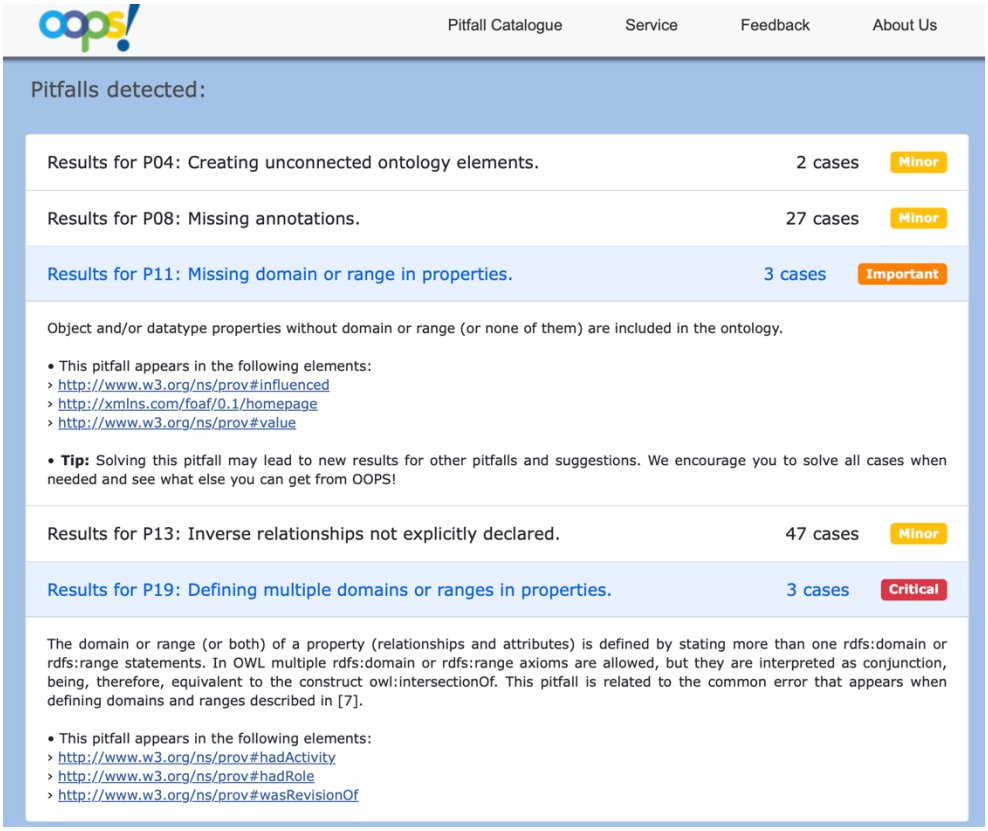
Back Submit Clear form

F.5: OOPS! and FOOPS! Results screenshot

OOPS:



FOOPS:



Appendix G: Study 4 Materials

G.1: Ethics application approval letter:

22/01/2026

Title: Evaluating the Impact of Metadata on expert Decision-Making for Mapping Reuse

REAMs No: 5226

Dear Sarah Alzahrani,

Your application has been reviewed by the relevant Research Ethics Committee (REC) and we are pleased to inform you that the above project has been approved.

Please note that any documents submitted for GDPR purposes in association with your ethics application are approved by the REC from an ethical perspective only (GDPR compliance is assumed, as per your uploaded letter of approval from the DPO if applicable).

It is the responsibility of the researcher/research team to ensure that, if applicable, all aspects of the study are executed in compliance with the General Data Protection Regulation (GDPR), the Health Research Regulations, Data Protection Act 2018, particularly regarding the storage and destruction of data arising from the research, and adhere to any relevant Health & Safety regulations.

Please note the reporting requirements outlined, in particular the need for:

- An immediate report of any serious or unexpected adverse event should be sent immediately to your REC via email. The Procedures for Reporting Adverse Events are outlined in the application. The Adverse Event form are available here.
- Any other unforeseen events should also be reported using this form and sent to your REC via email. The Adverse Event form are available here.
- An end of project report is due one year from the end of project date. The form is available here. This should be sent to your REC via email.
- If the data collection period is longer than 1 year, an annual report should be submitted. The form is available here. This should be sent to your REC via email.

You, and your research team (where applicable), are responsible for conducting the research strictly in the manner outlined in your application and supporting documentation, and in compliance with it. Proposed changes or amendments to the approved study must be submitted to the REC for approval in advance of any changes or amendments being implemented. In the event that the changes or amendments are significant, the REC may require a new submission. This ethical approval expires with the end dates of the project as stated in the application.

If you have any queries regarding this decision, please contact the Chair of your REC.

We wish you all the very best with your research project.

Kind Regards,

The REC

NB. If your study is to be conducted outside TCD, e.g. in a hospital site, please note that Trinity College REC approval does not of itself entitle you to conduct research in an external site. In this regard, the onus is upon you - the researcher, to ensure that all permissions have been obtained, and that any other requirements of the study site have been satisfied

NB. If your study involves a trial of medicinal products or medical devices, apps on humans you must provide a copy of your Clinical Trial Sponsorship from TCD - indicating that you have the appropriate trial insurance and sponsor oversight"

G.2: Invitation Email

Dear XXXXXXX

I hope you are doing well.

I am conducting a short expert study as part of my PhD research at Trinity College Dublin, under the supervision of Professor Declan O'Sullivan. This study investigates how metadata documentation can support the discovery of mapping artefacts for the purpose of reuse across different mapping types (e.g., RML mappings, ontology alignments, and interlinking rules).

The study presents three short scenarios, each involving a different type of mapping artefact. For each scenario, two alternative metadata representations are provided, and you are asked to evaluate how well they support reuse decisions. The focus is on whether the metadata model is sufficient to help you identify and select a mapping for reuse.

The study takes approximately 15–20 minutes to complete and does not require running any tools or inspecting mappings in detail.

If you are willing to participate, you can access the study here:
https://metasemap-1-sarah-alzahrani.replit.app/reuse

Your feedback will help in understanding how to design metadata that better supports mapping discovery and reuse.

This is the final experiment in my PhD evaluation, and your expert feedback would be extremely valuable. I would also like to sincerely thank you for taking part in my previous studies, I really appreciate your support.

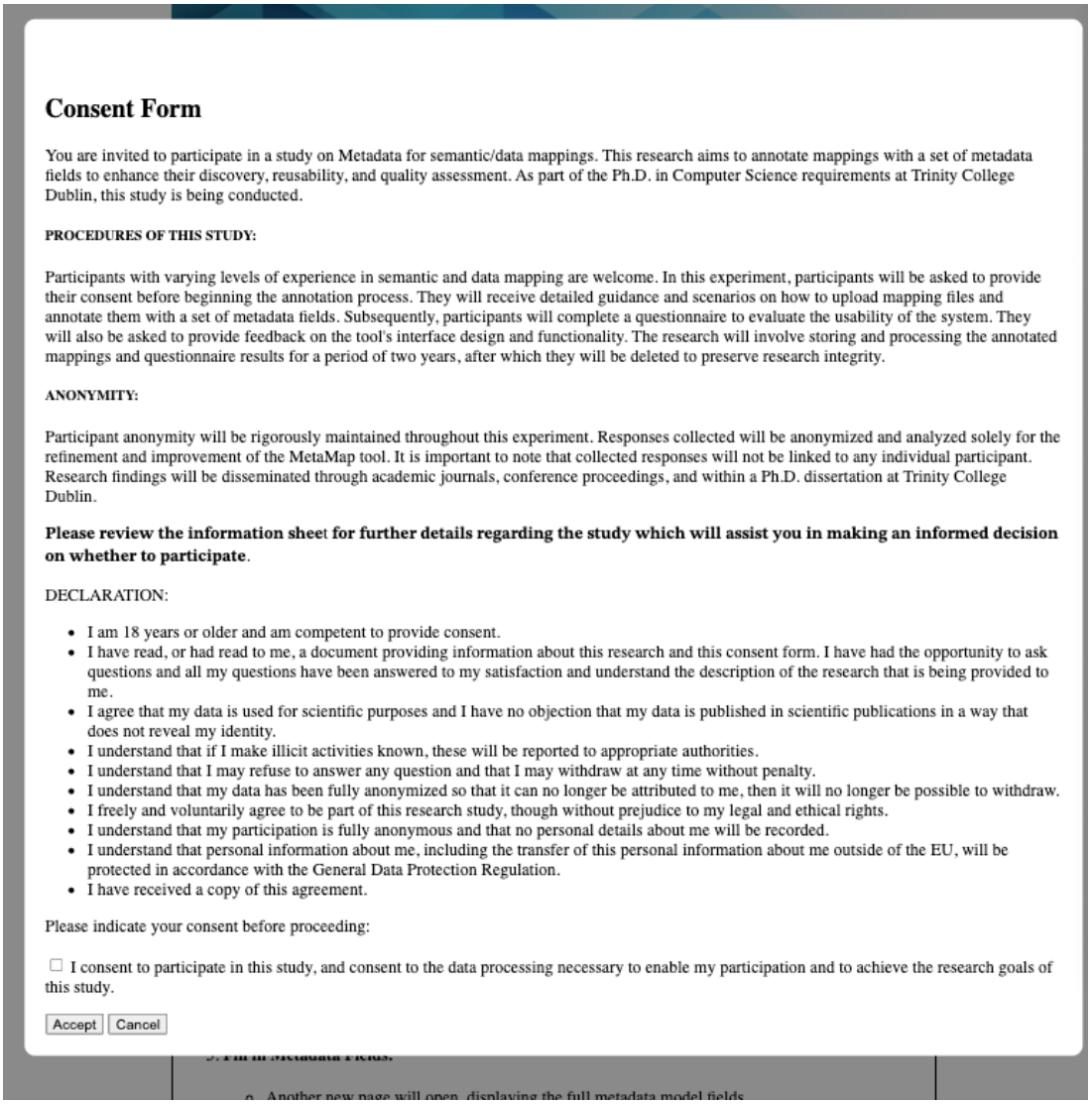
Participation is totally voluntary, and responses are anonymous and used solely for academic research purposes.

Please feel free to contact me if you have any questions.

Many thanks again for your time and support.

Best regards,
Sarah Alzahrani
PhD Researcher
Trinity College Dublin

G.3: Consent form



Consent Form

You are invited to participate in a study on Metadata for semantic/data mappings. This research aims to annotate mappings with a set of metadata fields to enhance their discovery, reusability, and quality assessment. As part of the Ph.D. in Computer Science requirements at Trinity College Dublin, this study is being conducted.

PROCEDURES OF THIS STUDY:

Participants with varying levels of experience in semantic and data mapping are welcome. In this experiment, participants will be asked to provide their consent before beginning the annotation process. They will receive detailed guidance and scenarios on how to upload mapping files and annotate them with a set of metadata fields. Subsequently, participants will complete a questionnaire to evaluate the usability of the system. They will also be asked to provide feedback on the tool's interface design and functionality. The research will involve storing and processing the annotated mappings and questionnaire results for a period of two years, after which they will be deleted to preserve research integrity.

ANONYMITY:

Participant anonymity will be rigorously maintained throughout this experiment. Responses collected will be anonymized and analyzed solely for the refinement and improvement of the MetaMap tool. It is important to note that collected responses will not be linked to any individual participant. Research findings will be disseminated through academic journals, conference proceedings, and within a Ph.D. dissertation at Trinity College Dublin.

Please review the information sheet for further details regarding the study which will assist you in making an informed decision on whether to participate.

DECLARATION:

- I am 18 years or older and am competent to provide consent.
- I have read, or had read to me, a document providing information about this research and this consent form. I have had the opportunity to ask questions and all my questions have been answered to my satisfaction and understand the description of the research that is being provided to me.
- I agree that my data is used for scientific purposes and I have no objection that my data is published in scientific publications in a way that does not reveal my identity.
- I understand that if I make illicit activities known, these will be reported to appropriate authorities.
- I understand that I may refuse to answer any question and that I may withdraw at any time without penalty.
- I understand that my data has been fully anonymized so that it can no longer be attributed to me, then it will no longer be possible to withdraw.
- I freely and voluntarily agree to be part of this research study, though without prejudice to my legal and ethical rights.
- I understand that my participation is fully anonymous and that no personal details about me will be recorded.
- I understand that personal information about me, including the transfer of this personal information about me outside of the EU, will be protected in accordance with the General Data Protection Regulation.
- I have received a copy of this agreement.

Please indicate your consent before proceeding:

☐ I consent to participate in this study, and consent to the data processing necessary to enable my participation and to achieve the research goals of this study.

Accept Cancel

G.4: Scenario descriptions and questionnaire

The survey instrument was administered using MetaSEMAP and the experiment accessible via the following link: https://metasemap-1-sarah-alzahrani.replit.app/

Expert Study on Metadata for Mapping Reuse

This study investigates how metadata can help experts decide whether an existing **mapping artefact** can be reused in a new project. The research forms part of a PhD in Computer Science at **Trinity College Dublin**.

In this study, a **mapping artefact** means a technical specification used to transform, align, or connect data. Examples include an **interlinking rule**, an **RML mapping**, and an **ontology alignment**.

Goal of the experiment: The goal is to evaluate whether different metadata models provide enough context for an expert to understand a mapping artefact, judge its trustworthiness, and decide whether it could be reused as a starting point in another setting.

About the two metadata models:

Representation A: FAIR-IMPACT metadata model
This model provides a concise description of the mapping artefact itself, such as its identifier, source and target resources, creator, tool, context, and evaluation information when available.

Representation B: MMV lifecycle-based metadata model
This model describes the mapping artefact across its lifecycle. In addition to basic artefact information, it may include details about the purpose of the mapping, assumptions, design decisions, technical requirements, risks, testing, maintenance, and stakeholders.

What you will do

1. **Review three reuse scenarios.**
   Each scenario presents one mapping artefact from a different type of task:
   - **Scenario 1:** Interlinking rule
   - **Scenario 2:** RML mapping
   - **Scenario 3:** Ontology alignment
2. **Compare two metadata representations.**
   In each scenario, the same mapping artefact is described using two metadata models:
   - **Representation A:** FAIR-IMPACT metadata model
   - **Representation B:** MMV lifecycle-based metadata model
3. **Make a reuse decision.**
   Your task is not to execute the mapping or inspect every rule in detail. Instead, please imagine that you have discovered this artefact while searching for reusable resources, and decide whether the metadata gives you enough information to reuse it as a starting point.
4. **Complete a short questionnaire after each scenario.**
   You will be asked whether you would consider the artefact reusable, which representation better supported your decision, which one was easier to understand, and what information was most useful or missing.

Important: You are not being asked to judge the full technical correctness of the artefact itself. You are being asked whether the metadata description gives you enough information to make an informed reuse decision.

Please note: Please base your responses only on the information shown in the study pages. External searching or additional resources are not required.

Estimated duration: approximately **15–20 minutes** in total.

For questions about the study, please contact salzahra@tcd.ie .

Proceed

Scenario 1: Reusing an Interlinking Rule

Artefact: **A Silk link specification** that generates **owl:sameAs** links between two datasets.

Scenario:
Imagine that you are working on a knowledge graph project and need to connect place entities in your dataset to entities in a well-known external source. You find an existing Silk interlinking rule that links cities in DBpedia and GeoNames.

Your task:
Review the artefact and the two metadata representations shown below. Then decide whether, based on the metadata alone, you would consider this artefact reusable as a starting point for a similar interlinking task in a new project.

What you are evaluating:
You are not being asked to judge the full technical correctness of the interlinking rule. You are being asked whether the metadata description gives you enough information to make an informed reuse decision.

Please note:
You do not need to inspect every line of the rule in detail. Please base your decision only on the metadata shown on this page and assume that detailed technical inspection is not available at this stage.

▼ **Silk link specification (excerpt)**

```
<Silk> <Prefixes> <Prefix id="dbpedia" namespace="http://dbpedia.org/resource/" /> <Prefix id="geonames"
namespace="http://sws.geonames.org/" /> <Prefix id="owl" namespace="http://www.w3.org/2002/07/owl#" /> <Prefix id="rdfs"
namespace="http://www.w3.org/2000/01/rdf-schema#" /> <Prefix id="wgs84_pos" namespace="http://www.w3.org/2003/01/geo/wgs84_pos#" />
</Prefixes> <DataSources> <DataSource id="dbpedia"> <EndpointURI>http://dbpedia.org/sparql</EndpointURI> </DataSource> <DataSource
id="geonames"> <EndpointURI>http://sws.geonames.org/sparql</EndpointURI> </DataSource> </DataSources> <Interlinks> <Interlink id="cities">
<LinkType>owl:sameAs</LinkType> <SourceDataset dataSource="dbpedia" var="a"> <RestrictTo>?a rdf:type dbpedia:City</RestrictTo>
</SourceDataset> <TargetDataset dataSource="geonames" var="b"> <RestrictTo>?b rdf:type geonames:Feature</RestrictTo> </TargetDataset>
<LinkCondition> <AVG> <Compare metric="jaro" weight="0.6"> <Param name="str1" path="?a/rdfs:label" /> <Param name="str2" path="?b/gn:name"
/> </Compare> <Compare metric="numSimilarity" optional="1" weight="0.7"> <Param name="num1" path="?a/dbpedia:populationTotal" /> <Param
name="num2" path="?b/gn:population" /> </Compare> <Compare metric="numSimilarity" optional="1" weight="0.7"> <Param name="num1" path="?
a/wgs84_pos:lat" /> <Param name="num2" path="?b/wgs84_pos:lat" /> </Compare> <Compare metric="numSimilarity" optional="1" weight="0.7">
<Param name="num1" path="?a/wgs84_pos:long" /> <Param name="num2" path="?b/wgs84_pos:long" /> </Compare> </AVG> </LinkCondition>
<Thresholds accept="0.9" verify="0.7" /> <Limit max="1" method="metric_value" /> <Output acceptedLinks="accepted_links.n3"
verifyLinks="verify_links.n3" mode="truncate" /> </Interlink> </Interlinks> </Silk>
```

(For this study, please use only the information provided on this page.)

Show metadata (Representation A: FAIR-IMPACT metadata model)

Values marked *Not provided* were not available from the scenario sources. This representation focuses on concise artefact-level metadata.

Representation A: FAIR-IMPACT metadata model Scenario 1

Metadata element	Scenario 1 value
Mapping-set (crosswalk) level Mapping-level	
mapping_set_id	`http://example.org/silk/interlink/cities/dbpedia-geonames`
source & target semantic artefact IDs	Source endpoint: `http://dbpedia.org/sparql` Target endpoint: `http://sws.geonames.org/sparql`
source & target semantic artefact names	Source: DBpedia (cities) Target: GeoNames (features / places)
version of the semantic artefacts	*Not provided*
specific types of semantic artefacts mapped	Place entities (cities / populated places) across two public linked datasets
mapping context	Entity interlinking for place records: the rule computes similarity using labels and optional numeric comparisons (population, latitude, longitude) and produces `owl:sameAs` links.
version (mapping set)	*Not provided*
date of creation of this version	*Not provided*
license	*Not provided*
author	*Not provided*
creator	*Not provided*
reviewer	*Not provided*
software used to create the mapping	**Silk** framework (link specification language with `<Interlink>` and `<LinkCondition>`)
validation / evaluation (mapping-level summary)	Thresholds: accept `0.9`, verify `0.7`; limit max `1` using `metric_value`; outputs: `accepted_links.n3` and `verify_links.n3` (mode truncate).

Show metadata (Representation B: MMV lifecycle-based metadata model)

Values marked *Not provided* were not available from the scenario sources. This representation provides lifecycle-oriented metadata, including purpose, assumptions, design, technical requirements, testing, and risks.

Representation B: MMV lifecycle-based metadata model Scenario 1

MMV field (phase-based)	Scenario 1 value
Project & Artefact	
mmv:MappingProject	`ex:Silk_DBpedia_GeoNames_Cities_Project`
mmv:MappingArtifact	Silk link specification that generates `owl:sameAs` links for cities between DBpedia and GeoNames
mmv:hasIdentifier	`http://example.org/silk/interlink/cities/dbpedia-geonames`
mmv:hasDescription	Interlink rule uses an AVG aggregation over similarity comparisons (label similarity with Jaro; optional numeric similarity for population and coordinates), with accept/verify thresholds and N3 output files.
mmv:version	*Not provided*
mmv:hasVersionDateTime	*Not provided*
Analysis phase	
mmv:hasPurpose	Interlink equivalent city/place entities across DBpedia and GeoNames by generating `owl:sameAs` links to support integration and enrichment.
mmv:hasMappingDomain	Geospatial / places (cities)
mmv:hasAssumption	City identity can be approximated using labels plus optional numeric evidence such as population and geo-coordinates; thresholds separate accepted links from links requiring verification.
mmv:hasDomainAssumption	*Not provided*
mmv:hasRequirement	Generate candidate identity links between DBpedia City resources and GeoNames Feature resources.
mmv:hasTechnicalRequirement	SPARQL access to both endpoints (`dbpedia` and `geonames` data sources) and ability to output N3 link files.
mmv:hasRiskOrIssue	Ambiguous place names; incomplete population/coordinate fields; endpoint availability or rate limits; possible false positives when different cities share similar labels.
mmv:hasInputName	DBpedia cities; GeoNames place features
mmv:hasInputDescription	*Not provided*
mmv:hasInputFormat	RDF accessed via SPARQL endpoints
mmv:hasInputType	Linked Data datasets (SPARQL)
mmv:hasInputSource	`http://dbpedia.org/sparql` and `http://sws.geonames.org/sparql`
mmv:hasInputCreator	*Not provided*

Questions (Scenario 1)

1) Based on the metadata shown, would you consider this artefact reusable as a starting point in a new project?

Representation A: ○ Yes ○ Possibly, but with reservations ○ No

Representation B: ○ Yes ○ Possibly, but with reservations ○ No

If you selected "Possibly, but with reservations", please briefly explain your reservations (for either representation):

(Optional)

2) Which representation better supported your reuse decision?

- ○ Representation A
- ○ Representation B
- ○ Both equally
- ○ Neither

3) Which representation was easier to understand?

- ○ Representation A
- ○ Representation B
- ○ Both equally
- ○ Neither

4) Which representation gave you more confidence in your decision?

- ○ Representation A
- ○ Representation B
- ○ Both equally
- ○ Neither

5) Do you think the two metadata models complement each other?

- ○ Yes, clearly
- ○ Somewhat
- ○ No, they mostly overlap
- ○ No, they are not compatible
- ○ Not sure

6) What information was most useful for making your reuse decision?

(Optional)

7) What important information was missing?

(Optional)

8) Any additional comments?

(Optional)

Submit and continue to Scenario 2 →

Scenario 2: Reusing an RML Mapping

Artefact: **A YARRRML/RML mapping** used to generate an RDF knowledge graph.

Scenario:
Imagine that you are building a knowledge graph for a new railway data integration project. You find an existing mapping from the European Union Agency for Railways (ERA), where YARRRML rules are converted to RML and executed with RMLMapper to transform source data into RDF.

Your task:
Review the artefact and the two metadata representations shown below. Then decide whether, based on the metadata alone, you would consider this mapping reusable as a starting point for a similar mapping task in a new setting.

What you are evaluating:
You are not being asked to judge the full technical correctness of the mapping itself. You are being asked whether the metadata description gives you enough information to make an informed reuse decision.

Please note:
You do not need to inspect every mapping rule in detail. Please base your decision only on the metadata shown on this page and assume that detailed technical inspection is not available at this stage.

▼ **Mapping file (excerpt)**

```
prefixes: grel: "http://users.ugent.be/~bjdmeest/function/grel.ttl#" eu-country:
"http://publications.europa.eu/resource/authority/country/" rdfs: "http://www.w3.org/2000/01/rdf-schema#" era:
"http://data.europa.eu/949/" xsd: "http://www.w3.org/2001/XMLSchema#" variables: connection: access: &host ${RINF_HOST} type: &typeMSSQL
mssql credentials: &credentials username: ${RINF_USER} password: ${RINF_PWD} queryFormulation: &queryFormulation sql2008 mappings: line-
national-ids: sources: - access: *host type: *typeMSSQL credentials: *credentials queryFormulation: *queryFormulation query: SELECT
DISTINCT UICLines.UICCode, MemberStates.Code FROM ... s: http://data.europa.eu/949/functionalInfrastructure/nationalLines/$(UICCode) po: -
[a, era:NationalRailwayLine] - [rdfs:label, $(UICCode)] - p: era:inCountry o: - mapping: eu-country condition: function: equal parameters:
- [str1, $(Code)] - [str2, $(code)] graph: ${NAMED_GRAPH} eu-country: sources: - ["data/countries.csv~csv"] s: $(eu)
```

(Please use only the information provided on this page for your decision.)

Show metadata (Representation A: FAIR-IMPACT metadata model)

Values marked *Not provided* were not available from the scenario sources. This representation focuses on concise artefact-level metadata.

Metadata element	Scenario 2 value
Mapping-set (crosswalk) level Mapping-level	
mapping_set_id	*Not provided*
source & target semantic artefact IDs	Source: ERA base registry data sources (databases, CSV files, XML files, etc.) Target: ERA Knowledge Graph (generated RDF knowledge graph)
source & target semantic artefact names	Source: Railway registry datasets used by ERA (e.g., base registry databases and lookup files) Target: ERA Knowledge Graph (uses ERA vocabulary namespace, e.g., `http://data.europa.eu/949/`)
version of the semantic artefacts	*Not provided*
specific types of semantic artefacts mapped	Railway registry sources → RDF knowledge graph (ERA vocabulary namespace visible in excerpt: `http://data.europa.eu/949/`)
mapping context	ERA use case: generates a knowledge graph to foster interoperability across base registry databases populated by EU member states and different actors from the railway domain.
version (mapping set)	*Not provided*
date of creation of this version	*Not provided*
license	*Not provided*
author	*Not provided*
creator	Mapping rules defined in **YARRRML** (automatically converted to **RML**), as stated in the use case.
reviewer	*Not provided*
software used to create the mapping	Use case states: **YARRRML** rules (converted to **RML**) executed with **RMLMapper**.
validation / evaluation (mapping-level summary)	*Not provided*

Show metadata (Representation B: MMV lifecycle-based metadata model)

Values marked *Not provided* were not available from the scenario sources. This representation provides lifecycle-oriented metadata, including purpose, assumptions, design, technical requirements, testing, and risks.

MMV field (phase-based)	Scenario 2 value
Project & Artefact	
mmv:MappingProject	`ex:ERA_RML_UseCase_Project`
mmv:MappingArtifact	YARRRML mappings (converted to RML) used to generate the ERA Knowledge Graph
mmv:hasIdentifier	*Not provided*
mmv:hasDescription	ERA use case: YARRRML (converted to RML) defines KG generation from different sources (databases, CSV, XML, etc.); RMLMapper executes the rules to generate the knowledge graph.
mmv:version	*Not provided*
mmv:hasVersionDateTime	*Not provided*
Analysis phase	
mmv:hasPurpose	Generate a knowledge graph to foster interoperability across railway registry databases.
mmv:hasMappingDomain	Railway domain interoperability
mmv:hasAssumption	Mappings rely on identifiers and codes across sources (e.g., the excerpt shows UIC codes, member state codes, and a country lookup mapping).
mmv:hasDomainAssumption	*Not provided*
mmv:hasRequirement	Configure data access for required sources and execute mappings to generate RDF output.
mmv:hasTechnicalRequirement	Execution environment: either Docker-based run or, without Docker, Node.js (≥ v12) and Java JRE (≥ v11), plus environment variables and connection parameters.
mmv:hasRiskOrIssue	Reuse for a new country may require adapting source schemas, joins, code lists, and connection settings such as host, username, password, and named graph configuration.
mmv:hasInputName	Railway registry databases; CSV lookup files; the use case also mentions XML files

mmv:hasInputDescription	*Not provided*
mmv:hasInputFormat	Databases; CSV files; XML files
mmv:hasInputType	Registry datasets
mmv:hasInputSource	ERA use case documentation and mappings repository (as referenced in the scenario sources)
mmv:hasInputCreator	*Not provided*
Design phase	
mmv:hasDesignDecision	Use YARRRML/YAML converted to RML. Map outputs to the ERA vocabulary (namespace visible in the excerpt: `http://data.europa.eu/949/`). Use lookup resources such as EU country authority data.
mmv:hasJustification	Use case indicates YARRRML is used to define knowledge graph generation rules and is automatically converted to RML for execution.
mmv:hasPlannedQualityMetric	*Not provided*
Development phase	
mqv:hasMappingType	RML mapping (semantic uplift)
mqv:hasMappingFormat	YARRRML/YAML (converted to RML)
mqv:hasMappingMethod	Declarative mapping rules
mmv:hasMappingAlgorithm	*Not provided*
mmv:hasTool	RMLMapper (as stated in the use case)

Testing phase	
mmv:hasTestingType	*Not provided*
mmv:hasTestingResult	*Not provided*
mmv:hasQualityMeasurement	*Not provided*
Maintenance phase	
mmv:hasPublisherName	European Union Agency for Railways (since 2020) — RML.io use case
mmv:hasPublisherSource	`https://rml.io/usage/`
Stakeholders	
mmv:hasStakeholder	*Not provided*
mmv:hasRole	*Not provided*
mmv:hasBackground	*Not provided*

Questions (Scenario 2)

1) Based on the metadata shown, would you consider this artefact reusable as a starting point in a new project?

Representation A: ○ Yes ○ Possibly, but with reservations ○ No

Representation B: ○ Yes ○ Possibly, but with reservations ○ No

If you selected "Possibly, but with reservations", please briefly explain your reservations (for either representation):

(Optional)

2) Which representation better supported your reuse decision?

- ○ Representation A
- ○ Representation B
- ○ Both equally
- ○ Neither

3) Which representation was easier to understand?

- ○ Representation A
- ○ Representation B
- ○ Both equally
- ○ Neither

4) Which representation gave you more confidence in your decision?

- ○ Representation A
- ○ Representation B
- ○ Both equally
- ○ Neither

5) Do you think the two metadata models complement each other?

- ○ Yes, clearly
- ○ Somewhat
- ○ No, they mostly overlap
- ○ No, they are not compatible
- ○ Not sure

6) What information was most useful for making your reuse decision?

(Optional)

7) What important information was missing?

(Optional)

8) Any additional comments?

(Optional)

Submit and continue to Scenario 3 →

After submitting, you will be taken directly to Scenario 3.

Scenario 3: Reusing an Ontology Alignment

Artefact: **An ontology alignment** between **mouse.owl** and **human.owl**.

Scenario:
Imagine that you are working on a biomedical knowledge graph project and need an existing alignment as a starting point for a similar anatomy alignment task. You find an ontology alignment produced in the OAEI 2024 Anatomy track, where the goal is to align concepts between the Adult Mouse Anatomy ontology (mouse.owl) and a human anatomy ontology (human.owl) so that equivalent anatomical structures across the two species can be related.

Your task:
Review the artefact and the two metadata representations shown below. Then decide whether, based on the metadata alone, you would consider this alignment reusable as a starting point for a similar task in a new project.

What you are evaluating:
You are not being asked to judge the full technical correctness of the alignment itself. You are being asked whether the metadata description gives you enough information to make an informed reuse decision.

Please note:
You do not need to inspect individual correspondences in detail. Please base your decision only on the metadata shown on this page and assume that detailed technical inspection is not available at this stage.

▼ **Alignment file (excerpt)**

```
<Alignment> <onto1>http://mouse.owl</onto1> <onto2>http://human.owl</onto2> <map> <Cell> <entity1
rdf:resource="http://mouse.owl#MA_0002668"/> <entity2 rdf:resource="http://human.owl#NCI_C32827"/> <relation>=</relation> <measure
rdf:datatype="http://www.w3.org/2001/XMLSchema#float">0.7</measure> </Cell> </map> <map> <Cell> <entity1
rdf:resource="http://mouse.owl#MA_0000378"/> <entity2 rdf:resource="http://human.owl#NCI_C12416"/> <relation>=</relation> <measure
rdf:datatype="http://www.w3.org/2001/XMLSchema#float">0.94</measure> </Cell> </map> <map> <Cell> <entity1
rdf:resource="http://mouse.owl#MA_0002492"/> <entity2 rdf:resource="http://human.owl#NCI_C12331"/> <relation>=</relation> <measure
rdf:datatype="http://www.w3.org/2001/XMLSchema#float">0.8</measure> </Cell> </map> </Alignment>
```

(Please use only the information provided on this page for your decision.)

▼ **Show metadata (Representation A: FAIR-IMPACT metadata model)**

Values marked *Not provided* were not available from the scenario source excerpt. This representation focuses on concise artefact-level metadata.

Metadata element	Scenario 3 value
Mapping-set (crosswalk) level Mapping-level	
mapping_set_id	`http://example.org/oaei/2024/anatomy/logmap`
source & target semantic artefact IDs	Source: `http://mouse.owl` Target: `http://human.owl`
source & target semantic artefact names	Source: mouse.owl (Adult Mouse Anatomy ontology) Target: human.owl (human anatomy ontology; NCI identifiers used in the alignment)
version of the semantic artefacts	*Not provided*
specific types of semantic artefacts mapped	Large biomedical anatomy ontologies (mouse anatomy ↔ human anatomy)
mapping context	OAEI 2024 Anatomy track benchmark evaluation. Reported metrics include precision, recall, F-measure and recall+. The track notes that recall+ reflects non-trivial correspondences where labels are not identical after normalization.
version (mapping set)	OAEI 2024 (Anatomy track) — LogMap system output (standard setting)
date of creation of this version	*Not provided*
license	*Not provided*
author	*Not provided*
creator	Alignment generated by: **LogMap** (Developer or creator details: *Not provided*)
reviewer	*Not provided*
software used to create the mapping	Matcher: **LogMap** Track evaluation platform: MELT was used for most systems (ALIN and MDMapper used SEALS client, according to the track report).
validation / evaluation (mapping-level summary)	Track report: systems executed in standard setting; metrics computed were precision, recall, F-measure and recall+. Track notes include filtering trivial `oboInOwl#Synonym`-style correspondences and non-equivalence relations; coherence was checked with Pellet by merging ontologies and alignment. **LogMap (OAEI 2024 Anatomy results):** Runtime 12s; Size 1402; Precision 0.917; F-measure 0.881; Recall 0.848; Recall+ 0.602; Coherent: +

▼ **Show metadata (Representation B: MMV lifecycle-based metadata model)**

Values marked *Not provided* were not available from the scenario source excerpt. This representation provides lifecycle-oriented metadata, including purpose, assumptions, design, technical requirements, testing, and risks.

MMV field (phase-based)	Scenario 3 value
Project & Artefact	
mmv:MappingProject	`ex:OAEI2024_Anatomy_LogMap_Project`
mmv:MappingArtifact	Ontology alignment output (Alignment API / RDF), e.g., `LogMap.rdf` from OAEI 2024 Anatomy track raw results
mmv:hasIdentifier	`http://example.org/oaei/2024/anatomy/logmap`
mmv:hasDescription	Alignment between mouse.owl and human.owl from the OAEI 2024 Anatomy track, represented using the Alignment API in RDF. Correspondences use an equivalence relation and include confidence measures.
mmv:version	OAEI 2024 (Anatomy track) — LogMap system output (standard setting)
mmv:hasVersionDateTime	*Not provided*
Analysis phase	
mmv:hasPurpose	Provide a benchmark alignment artefact for ontology interoperability and support reuse decisions in new anatomy alignment settings.
mmv:hasMappingDomain	Biomedical anatomy / ontology interoperability
mmv:hasAssumption	Track report notes a mix of trivial and non-trivial correspondences; recall+ reflects detection of non-trivial matches where labels differ after normalization.
mmv:hasDomainAssumption	*Not provided*
mmv:hasRequirement	Generate an alignment for mouse.owl and human.owl under the OAEI 2024 Anatomy track standard setting.
mmv:hasTechnicalRequirement	Executable in the OAEI evaluation environment; track report indicates MELT was used for most systems, while ALIN and MDMapper used SEALS.
mmv:hasRiskOrIssue	Reuse in a different setting may be affected by ontology drift, domain differences, and the proportion of non-trivial correspondences.
mmv:hasInputName	mouse.owl; human.owl
mmv:hasInputDescription	*Not provided*
mmv:hasInputFormat	*Not provided*
mmv:hasInputType	Ontology files (biomedical anatomy ontologies)

mmv:hasInputSource	OAEI 2024 Anatomy track (results page referenced in this study's scenario source)
mmv:hasInputCreator	*Not provided*
Design phase	
mmv:hasDesignDecision	Use Alignment API / RDF representation and represent correspondences with equivalence relation and confidence measure.
mmv:hasJustification	Alignment API is commonly used in OAEI outputs and supports interoperability across tools and evaluations.
mmv:hasPlannedQualityMetric	Precision, Recall, F-measure, Recall+, plus coherence checking
Development phase	
mqv:hasMappingType	Ontology alignment
mqv:hasMappingFormat	Alignment API / RDF
mqv:hasMappingMethod	Automatic ontology matching system execution (standard setting)
mmv:hasMappingAlgorithm	*Not provided*
mmv:hasTool	LogMap (matcher); track evaluation used MELT for most systems
Testing phase	
mmv:hasTestingType	Benchmark evaluation (OAEI 2024 Anatomy track). Metrics computed; track report also describes filtering trivial and non-equivalence correspondences and coherence checking with Pellet.
mmv:hasTestingResult	**LogMap (OAEI 2024 Anatomy track):** Runtime: 12s; Alignment size: 1402; Precision: 0.917; Recall: 0.848; F-measure: 0.881; Recall+: 0.602; Coherent alignment: +
mmv:hasQualityMeasurement	Precision, Recall, F-measure, Recall+, Coherence
Maintenance phase	
mmv:hasPublisherName	OAEI 2024 Anatomy track
mmv:hasPublisherSource	Results page referenced in scenario source: `http://oaei.ontologymatching.org/2024/results/anatomy/`
Stakeholders	
mmv:hasStakeholder	Track organizers: Mina Abd Nikooie Pour; Huanyu Li; Ying Li; Patrick Lambrix
mmv:hasRole	Track organizers / evaluators
mmv:hasBackground	*Not provided*

Questions (Scenario 3)

1) Based on the metadata shown, would you consider this artefact reusable as a starting point in a new project?

Representation A: ○ Yes ○ Possibly, but with reservations ○ No

Representation B: ○ Yes ○ Possibly, but with reservations ○ No

If you selected "Possibly, but with reservations", please briefly explain your reservations (for either representation):

(Optional)

2) Which representation better supported your reuse decision?

- ○ Representation A
- ○ Representation B
- ○ Both equally
- ○ Neither

3) Which representation was easier to understand?

- ○ Representation A
- ○ Representation B
- ○ Both equally
- ○ Neither

4) Which representation gave you more confidence in your decision?

- ○ Representation A
- ○ Representation B
- ○ Both equally
- ○ Neither

5) Do you think the two metadata models complement each other?

- ○ Yes, clearly
- ○ Somewhat
- ○ No, they mostly overlap
- ○ No, they are not compatible
- ○ Not sure

6) What information was most useful for making your reuse decision?

(Optional)

7) What important information was missing?

(Optional)

8) Any additional comments?

(Optional)

Submit Scenario 3

After submitting, you will see a final "thank you" page.



Final Questions

Thank you.
You have now completed the three scenarios. These final questions ask about your overall preference after seeing all three mapping artefacts.

1) After completing all three scenarios, which metadata representation would you prefer to use in practice when documenting mapping artefacts?

- ○ Representation A (FAIR-IMPACT metadata model)
- ○ Representation B (MMV lifecycle-based metadata model)
- ○ Both together
- ○ No preference

2) Why?

Please briefly explain your choice.

3) Any final suggestions for improving metadata documentation for mapping reuse?

(Optional)

Submit Final Responses

Copyright 2026, Sarah Alzahrani salzahra@tcd.ie



Appendix H: Use Case: Metadata-Guided LLM Mapping Generation (Chapter 6) Materials

RML Mappings

S1A- Baseline prompt (no metadata)

```
1. Generate a RML mapping artefact for the following task.
2. Task description:
3. You are given a CSV dataset containing European railway infrastructure data with the following
   columns:
4. OperationalPointName, OperationalPointType, Latitude, Longitude, CountryCode,
   LineIdentification.
5. The dataset describes railway operational points such as stations, junctions, and signals in
   Europe.
6. Generate an RML mapping that converts this CSV into RDF.
7. Return only valid RML in Turtle syntax.
8. Return only the RML output with no extra explanation.
```

S1A- Baseline prompt result (no metadata)

```
1.  @prefix rr: <http://www.w3.org/ns/r2rml#> .
2.  @prefix rml: <http://semweb.mmlab.be/ns/rml#> .
3.  @prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
4.  @prefix ex: <http://example.org/> .
5.
6.  ex:RailwayOperationalPointsMapping a rml:Mapping ;
7.      rml:logicalSource [
8.          rml:source "path/to/your/dataset.csv" ;
9.          rml:referenceFormulation rml:CSV ;
10.     ] ;
11.     rr:subjectMap [
12.         rr:template "http://example.org/railwayOperationalPoint/{OperationalPointName}" ;
13.         rr:class ex:OperationalPoint ;
14.     ] ;
15.     rr:predicateObjectMap [
16.         rr:predicate ex:hasName ;
17.         rr:objectMap [
18.             rml:column "OperationalPointName" ;
19.         ] ;
20.     ] ;
21.     rr:predicateObjectMap [
22.         rr:predicate ex:hasType ;
23.         rr:objectMap [
24.             rml:column "OperationalPointType" ;
25.         ] ;
26.     ] ;
27.     rr:predicateObjectMap [
28.         rr:predicate ex:hasLocation ;
29.         rr:objectMap [
```

```
            rr:template "http://www.w3.org/2003/01/geo/wgs84_pos#latlong" ;
            rr:graphMap [
                rr:template
    "http://example.org/railwayOperationalPoint/{OperationalPointName}/location" ;
            ] ;
            rr:class ex:Location ;
            rr:predicateObjectMap [
                rr:predicate ex:latitude ;
                rr:objectMap [
                    rml:column "Latitude" ;
                    rr:datatype xsd:decimal ;
                ] ;
            ] ;
            rr:predicateObjectMap [
                rr:predicate ex:longitude ;
                rr:objectMap [
                    rml:column "Longitude" ;
                    rr:datatype xsd:decimal ;
                ] ;
            ] ;
        ] ;
    ] ;
    rr:predicateObjectMap [
        rr:predicate ex:locatedIn ;
        rr:objectMap [
            rml:column "CountryCode" ;
            rr:template "http://example.org/country/{CountryCode}" ;
        ] ;
    ] ;
    rr:predicateObjectMap [
        rr:predicate ex:hasLineIdentification ;
        rr:objectMap [
            rml:column "LineIdentification" ;
        ] ;
    ] .
```

S1A- Guided prompt (with metadata)

```
Generate a RML mapping artefact for the following task.

Task description:
You are given a CSV dataset containing European railway
infrastructure data with the following columns:
OperationalPointName, OperationalPointType, Latitude,
Longitude, CountryCode, LineIdentification.

The dataset describes railway operational points such as
stations, junctions, and signals in Europe.
Generate an RML mapping that converts this CSV into RDF.
Return only valid RML in Turtle syntax.

Lifecycle metadata:
```

```
The following lifecycle metadata is drawn from the Mapping
Metadata Vocabulary (MMV: http://w3id.org/mmv/0.1). Use it
to guide your mapping generation decisions.

[mmv:AnalysisActivity]
mmv:hasPurpose: Publish European railway infrastructure as
Linked Data to support interoperability across national rail
networks and integration with the ERA Knowledge Graph.
mmv:hasMappingDomain: Railway / Transport Infrastructure
mmv:hasRequirement: All six columns must be represented.
Each operational point must receive a persistent URI
constructed from OperationalPointName.
mmv:hasAssumption: Latitude and Longitude are WGS84 decimal
degree coordinates. CountryCode is ISO 3166-1 alpha-2.
mmv:hasDomainAssumption: LineIdentification relates
operational points to railway lines and should be
represented as a typed relationship to a line resource,
not a literal string.
mmv:hasTechnicalRequirement: Valid RML in Turtle syntax
executable by RMLMapper.
mmv:hasRiskOrIssue: Using generic custom vocabulary reduces
interoperability with existing ERA Linked Data infrastructure.
mmv:hasInputName: railway_operational_points.csv
mmv:hasInputFormat: text/csv
[mmv:DesignActivity]
mmv:hasDesignDecision: Use established transport and
geographic vocabularies for infrastructure typing and
spatial coordinates. Represent CountryCode as a typed IRI
resource not a literal string.
mmv:hasJustification: ERA data must integrate with existing
European transport knowledge graphs. Standard vocabularies
enable cross-border data integration without custom mappings.
mmv:hasPlannedQualityMetric: All six columns mapped.
Coordinate values typed as xsd:decimal. Type classification
uses a controlled vocabulary term. Each row produces
exactly one resource.
[mqv:DevelopmentActivity]
mqv:hasMappingType: Uplift Mapping
mqv:hasMappingFormat: RML Turtle


After generating the RML mapping, append a metadata
annotation block in Turtle format using MMV properties:
Return only the RML output with no extra explanation.
```

S1A- Guided prompt result (with metadata)

```
@prefix rr: <http://www.w3.org/ns/r2rml#> .
@prefix rml: <http://semweb.mmlab.be/ns/rml#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
```

```
4.  @prefix ex: <http://example.org/resource/> .
5.  @prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
6.  @prefix schema: <http://schema.org/> .
7.
8.  <http://example.org/mapping/era-railway>
9.      a rml:Mapping ;
10.     rml:logicalSource [
11.         rml:source <file:///path/to/railway_operational_points.csv> ;
12.         rml:referenceFormulation rml:CSV ;
13.     ] ;
14.     rr:triplesMap [
15.         rr:subjectMap [
16.             rr:template "http://example.org/railway/{OperationalPointName}" ;
17.             rr:class schema:Place ;
18.         ] ;
19.         rr:predicateObjectMap [
20.             rr:predicate schema:name ;
21.             rr:objectMap [
22.                 rml:column "OperationalPointName" ;
23.             ] ;
24.         ] ;
25.         rr:predicateObjectMap [
26.             rr:predicate schema:additionalType ;
27.             rr:objectMap [
28.                 rml:column "OperationalPointType" ;
29.             ] ;
30.         ] ;
31.         rr:predicateObjectMap [
32.             rr:predicate geo:lat ;
33.             rr:objectMap [
34.                 rml:column "Latitude" ;
35.                 rml:datatype xsd:decimal ;
36.             ] ;
37.         ] ;
38.         rr:predicateObjectMap [
39.             rr:predicate geo:long ;
40.             rr:objectMap [
41.                 rml:column "Longitude" ;
42.                 rml:datatype xsd:decimal ;
43.             ] ;
44.         ] ;
45.         rr:predicateObjectMap [
46.             rr:predicate schema:addressCountry ;
47.             rr:objectMap [
48.                 rml:column "CountryCode" ;
49.                 rr:termType rml:IRI ;
50.                 rr:template "http://example.org/country/{CountryCode}" ;
51.             ] ;
52.         ] ;
53.         rr:predicateObjectMap [
54.             rr:predicate schema:relatedTo ;
55.             rr:objectMap [
```

```
                rml:column "LineIdentification" ;
                rr:termType rml:IRI ;
                rr:template "http://example.org/line/{LineIdentification}" ;
            ] ;
        ] ;
    ] .

<http://example.org/mapping/era-railway>
    a mmv:MappingArtifact ;
    mmv:hasPurpose "Publish European railway infrastructure as Linked Data for ERA Knowledge Graph integration." ;
    mmv:hasMappingDomain "Railway / Transport Infrastructure" ;
    mmv:hasDesignDecision "Use established transport and geographic vocabularies for infrastructure typing." ;
    mmv:hasJustification "Enables cross-border data integration with existing ERA knowledge graphs." ;
    mqv:hasMappingType "Uplift Mapping" ;
    mqv:hasMappingFormat "RML Turtle" ;
    mmv:hasRequirement "All six columns mapped. Persistent URIs from OperationalPointName." .
```

S1B- Baseline prompt (no metadata)

```
Generate a RML mapping artefact for the following task.
Task description:
You are given a CSV dataset in GTFS format with the
following columns from the stops.txt file:
stop_id, stop_name, stop_lat, stop_lon,
stop_url, location_type.
The dataset contains public transit stop records.
Generate an RML mapping that converts this CSV into RDF.
Return only valid RML in Turtle syntax.
Return only the RML output with no extra explanation.
```

S1B- Baseline prompt result (no metadata)

```
@prefix rr: <http://www.w3.org/ns/r2rml#> .
@prefix rml: <http://semweb.mmlab.be/ns/rml#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix ex: <http://example.org/> .

ex:GTFSMapping a rml:Mapping ;
    rml:logicalSource [
        rml:source "path/to/stops.txt" ;
        rml:referenceFormulation rml:CSV ;
    ] ;
    rr:subjectMap [
        rr:template "http://example.org/stop/{stop_id}" ;
        rr:class ex:Stop ;
    ] ;
    rr:predicateObjectMap [
```

```
        rr:predicate ex:stopName ;
        rr:objectMap [
            rml:column "stop_name" ;
        ] ;
    ] ;
    rr:predicateObjectMap [
        rr:predicate ex:stopLatitude ;
        rr:objectMap [
            rml:column "stop_lat" ;
            rr:datatype xsd:decimal ;
        ] ;
    ] ;
    rr:predicateObjectMap [
        rr:predicate ex:stopLongitude ;
        rr:objectMap [
            rml:column "stop_lon" ;
            rr:datatype xsd:decimal ;
        ] ;
    ] ;
    rr:predicateObjectMap [
        rr:predicate ex:stopUrl ;
        rr:objectMap [
            rml:column "stop_url" ;
        ] ;
    ] ;
    rr:predicateObjectMap [
        rr:predicate ex:locationType ;
        rr:objectMap [
            rml:column "location_type" ;
        ] ;
    ] .
```

S1B- Guided prompt (with metadata)

```
Generate a RML mapping artefact for the following task.

Task description:
You are given a CSV dataset in GTFS format with the
following columns from the stops.txt file:
stop_id, stop_name, stop_lat, stop_lon,
stop_url, location_type.

The dataset contains public transit stop records.
Generate an RML mapping that converts this CSV into RDF.
Return only valid RML in Turtle syntax.

Lifecycle metadata:
The following lifecycle metadata is drawn from the Mapping
Metadata Vocabulary (MMV: http://w3id.org/mmv/0.1). Use it
to guide your mapping generation decisions.
```

```
17.
18.  [mmv:AnalysisActivity]
19.  mmv:hasPurpose: Convert public transit stop data into RDF
20.  to enable integration with smart city knowledge graphs and
21.  multimodal journey planning applications.
22.  mmv:hasMappingDomain: Public Transport / Smart City
23.  mmv:hasRequirement: All six columns must be represented.
24.  Each stop must receive a persistent URI constructed
25.  from stop_id.
26.  mmv:hasAssumption: stop_lat and stop_lon are WGS84 decimal
27.  degree coordinates. stop_url is a resolvable web address
28.  and should be typed as an IRI not a literal.
29.  mmv:hasDomainAssumption: location_type uses GTFS encoding
30.  where 0 indicates a transit stop and 1 indicates a station.
31.  This distinction must be represented as a typed
32.  classification not a raw integer literal.
33.  mmv:hasTechnicalRequirement: Valid RML in Turtle syntax
34.  executable by RMLMapper.
35.  mmv:hasRiskOrIssue: Using generic custom vocabulary reduces
36.  interoperability with existing transport Linked Data and
37.  smart city knowledge graphs.
38.  mmv:hasInputName: stops.txt
39.  mmv:hasInputFormat: text/csv
40.  [mmv:DesignActivity]
41.  mmv:hasDesignDecision: Use an established public transport
42.  ontology for typing stops and stations. Use standard
43.  geographic properties for coordinates. Represent stop_url
44.  as a typed IRI. Represent location_type as a typed
45.  classification using a controlled vocabulary.
46.  mmv:hasJustification: GTFS data is widely used in transport
47.  applications. Established vocabularies ensure output
48.  integrates with existing transport Linked Data without
49.  additional transformation steps.
50.  mmv:hasPlannedQualityMetric: All six columns mapped.
51.  stop_lat and stop_lon typed as xsd:decimal. location_type
52.  represented as typed classification. stop_url typed as IRI.
53.  Each row produces exactly one resource.
54.  [mqv:DevelopmentActivity]
55.  mqv:hasMappingType: Uplift Mapping
56.  mqv:hasMappingFormat: RML Turtle
57.
58.  After generating the RML mapping, append a metadata
59.  annotation block in Turtle format using MMV properties:
60.  Return only the RML output with no extra explanation.
```

S1B- Guided prompt result (with metadata)

```
1.   @prefix rr: <http://www.w3.org/ns/r2rml#> .
2.   @prefix rml: <http://semweb.mmlab.be/ns/rml#> .
3.   @prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
4.   @prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
5.   @prefix transit: <http://vocab.org/transit/terms#> .
```

```
@prefix mmv: <http://w3id.org/mmv/0.1#> .
@prefix mqv: <https://alex-randles.github.io/MQV/#> .

<http://example.org/mapping/gtfs-stops>
    a rml:Mapping ;
    rml:logicalSource [
        rml:source "stops.txt" ;
        rml:referenceFormulation rml:CSV ;
    ] ;
    rr:triplesMap [
        rr:subjectMap [
            rr:template "http://example.org/stop/{stop_id}" ;
            rr:class transit:Stop ;
        ] ;
        rr:predicateObjectMap [
            rr:predicate transit:stopName ;
            rr:objectMap [
                rml:column "stop_name" ;
            ] ;
        ] ;
        rr:predicateObjectMap [
            rr:predicate geo:lat ;
            rr:objectMap [
                rml:column "stop_lat" ;
                rr:datatype xsd:decimal ;
            ] ;
        ] ;
        rr:predicateObjectMap [
            rr:predicate geo:long ;
            rr:objectMap [
                rml:column "stop_lon" ;
                rr:datatype xsd:decimal ;
            ] ;
        ] ;
        rr:predicateObjectMap [
            rr:predicate transit:stopUrl ;
            rr:objectMap [
                rml:column "stop_url" ;
                rr:datatype rml:IRI ;
            ] ;
        ] ;
        rr:predicateObjectMap [
            rr:predicate transit:locationType ;
            rr:objectMap [
                rml:column "location_type" ;
                rr:termType rr:Literal ;
                rr:datatype transit:LocationType ;
            ] ;
        ] ;
    ] .

<http://example.org/mapping/gtfs-stops>
```

```
58.      a mmv:MappingArtifact ;
59.      mmv:hasPurpose "Convert GTFS stop data to RDF for smart city knowledge graph integration."
     ;
60.      mmv:hasMappingDomain "Public Transport / Smart City" ;
61.      mmv:hasDesignDecision "Use established transport ontology for typing. Represent stop_url
     as IRI. Represent location_type as typed classification." ;
62.      mmv:hasJustification "Enables integration with existing transport Linked Data." ;
63.      mqv:hasMappingType "Uplift Mapping" ;
64.      mqv:hasMappingFormat "RML Turtle" ;
65.      mmv:hasRequirement "All six columns mapped. Persistent URIs from stop_id." .
```

Ontology Alignments

S2A- Ontology Alignment- Baseline prompt (no metadata)

```
1.  Generate a OntologyAlignment mapping artefact for the following task.
2.  Task description:
3.  You are given two ontologies describing anatomy.
4.  The source ontology describes mouse anatomy and the
5.  target ontology describes human anatomy.
6.  Generate an ontology alignment that identifies equivalent
7.  anatomical concepts between the two ontologies.
8.  Return the alignment in RDF/XML format using the
9.  Alignment API format with align:Alignment structure.
10. Each correspondence must include align:entity1,
11. align:entity2, align:measure, and align:relation.
12. Return only the OntologyAlignment output with no extra explanation.
```

S2A- Ontology Alignment- Baseline prompt result (no metadata)

```
1.  <align:Alignment xmlns:align="http://knowledgeweb.semanticweb.org/heterogeneity/alignment">
2.      <align:map>
3.          <align:cell>
4.              <align:entity1 rdf:resource="http://example.org/ontology/mouse#MouseHeart"/>
5.              <align:entity2 rdf:resource="http://example.org/ontology/human#HumanHeart"/>
6.              <align:measure>1.0</align:measure>
7.              <align:relation>equivalent</align:relation>
8.          </align:cell>
9.          <align:cell>
10.             <align:entity1 rdf:resource="http://example.org/ontology/mouse#MouseLung"/>
11.             <align:entity2 rdf:resource="http://example.org/ontology/human#HumanLung"/>
12.             <align:measure>1.0</align:measure>
13.             <align:relation>equivalent</align:relation>
14.         </align:cell>
15.         <align:cell>
16.             <align:entity1 rdf:resource="http://example.org/ontology/mouse#MouseLiver"/>
17.             <align:entity2 rdf:resource="http://example.org/ontology/human#HumanLiver"/>
18.             <align:measure>1.0</align:measure>
```

```
            <align:relation>equivalent</align:relation>
        </align:cell>
        <align:cell>
            <align:entity1 rdf:resource="http://example.org/ontology/mouse#MouseKidney"/>
            <align:entity2 rdf:resource="http://example.org/ontology/human#HumanKidney"/>
            <align:measure>1.0</align:measure>
            <align:relation>equivalent</align:relation>
        </align:cell>
        <align:cell>
            <align:entity1 rdf:resource="http://example.org/ontology/mouse#MouseBrain"/>
            <align:entity2 rdf:resource="http://example.org/ontology/human#HumanBrain"/>
            <align:measure>1.0</align:measure>
            <align:relation>equivalent</align:relation>
        </align:cell>
    </align:map>
</align:Alignment>
```

S2A- Ontology Alignment- Baseline prompt (metadata)

```
Generate a OntologyAlignment mapping artefact for the following task.

Task description:
You are given two ontologies describing anatomy.
The source ontology describes mouse anatomy and the
target ontology describes human anatomy.

Generate an ontology alignment that identifies equivalent
anatomical concepts between the two ontologies.

Return the alignment in RDF/XML format using the
Alignment API format with align:Alignment structure.
Each correspondence must include align:entity1,
align:entity2, align:measure, and align:relation.

Lifecycle metadata:
The following lifecycle metadata is drawn from the Mapping
Metadata Vocabulary (MMV: http://w3id.org/mmv/0.1). Use it
to guide your alignment generation decisions.

[mmv:AnalysisActivity]
mmv:hasPurpose: Support cross-species biomedical research
by identifying equivalent anatomical structures between
mouse and human anatomy ontologies to enable data
integration across species-specific datasets.
mmv:hasMappingDomain: Biomedical / Anatomy
mmv:hasRequirement: Each correspondence must reference
real URIs from the source and target ontologies.
Each cell must include entity1, entity2, measure,
and relation fields.
mmv:hasAssumption: Anatomical equivalence is based on
```

```
functional and structural homology not surface-level
label similarity. Some concepts may have no equivalent
in the target ontology.
mmv:hasDomainAssumption: The source ontology is the
Adult Mouse Anatomy ontology. Its classes use the
namespace http://mouse.owl# with MA_ identifiers
for example http://mouse.owl#MA_0000003 for the
alimentary system. The target ontology is the NCI
Thesaurus human anatomy subset. Its classes use
the namespace http://human.owl# with NCI_ identifiers
for example http://human.owl#NCI_C12378 for the
alimentary system.
mmv:hasTechnicalRequirement: Valid Alignment API
RDF/XML format. Relation must use = for equivalence
not the word equivalent.
mmv:hasRiskOrIssue: Fabricating URIs that do not
exist in the actual ontologies makes the alignment
impossible to evaluate and unusable in practice.
mmv:hasInputName: mouse.owl and human.owl
mmv:hasInputFormat: OWL/RDF
[mmv:DesignActivity]
mmv:hasDesignDecision: Use only URIs that exist in
the actual OAEI anatomy ontologies. Assign confidence
values between 0 and 1 reflecting degree of certainty
not a uniform 1.0 for all correspondences. Use =
as the relation for equivalence correspondences.
mmv:hasJustification: The alignment will be evaluated
against the published OAEI anatomy gold standard.
Fabricated URIs score zero against the gold standard.
Realistic confidence values enable threshold-based
filtering in downstream applications.
mmv:hasPlannedQualityMetric: Precision and Recall
against OAEI anatomy gold standard. All entity URIs
must resolve to real ontology terms. Confidence
values must vary to reflect certainty.
[mqv:DevelopmentActivity]
mqv:hasMappingType: Ontology Alignment
mqv:hasMappingFormat: Alignment API RDF/XML
After generating the alignment, append a metadata
block using MMV:
Return only the OntologyAlignment output with no extra explanation.
```

S2A- Ontology Alignment- Baseline prompt result (metadata)

```
<align:Alignment xmlns:align="http://knowledgeweb.semanticweb.org/heterogeneity/alignment">
    <align:entity1 rdf:resource="http://mouse.owl#MA_0000003"/>
    <align:entity2 rdf:resource="http://human.owl#NCI_C12378"/>
    <align:measure>0.9</align:measure>
    <align:relation>=</align:relation>

    <align:entity1 rdf:resource="http://mouse.owl#MA_0000004"/>
    <align:entity2 rdf:resource="http://human.owl#NCI_C12379"/>
```

```
      <align:measure>0.85</align:measure>
      <align:relation>=</align:relation>

      <align:entity1 rdf:resource="http://mouse.owl#MA_0000005"/>
      <align:entity2 rdf:resource="http://human.owl#NCI_C12380"/>
      <align:measure>0.8</align:measure>
      <align:relation>=</align:relation>

      <align:entity1 rdf:resource="http://mouse.owl#MA_0000006"/>
      <align:entity2 rdf:resource="http://human.owl#NCI_C12381"/>
      <align:measure>0.75</align:measure>
      <align:relation>=</align:relation>
</align:Alignment>

@prefix mmv: <http://w3id.org/mmv/0.1#> .
@prefix mqv: <https://alex-randles.github.io/MQV/#> .

<http://example.org/alignment/oaei-anatomy>
      a mmv:MappingArtifact ;
      mmv:hasPurpose "Cross-species anatomy alignment for biomedical data integration." ;
      mmv:hasMappingDomain "Biomedical / Anatomy" ;
      mmv:hasDesignDecision "Use real OAEI ontology URIs. Use = for equivalence. Vary confidence
    values." ;
      mmv:hasJustification "Enables evaluation against OAEI gold standard." ;
      mqv:hasMappingType "Ontology Alignment" ;
      mqv:hasMappingFormat "Alignment API RDF/XML" .
```

S2B- Ontology Alignment- Baseline prompt (no metadata)

```
Generate a OntologyAlignment mapping artefact for the following task.

Task description:
You are given two ontologies describing academic
conference organisation. The source ontology is cmt
and the target ontology is the conference ontology.
Both describe concepts related to academic events
including papers, authors, reviews, and programme
committees.

Generate an ontology alignment that identifies
equivalent concepts between the two ontologies.

Return the alignment in RDF/XML format using the
Alignment API format with align:Alignment structure.
Each correspondence must use align:Cell with
align:entity1, align:entity2, align:measure,
and align:relation.

Return only the OntologyAlignment output with no extra explanation.
```

S2B- Ontology Alignment- Baseline prompt result (no metadata)

```
<align:Alignment xmlns:align="http://knowledgeweb.semanticweb.org/heterogeneity/alignment">
    <align:cell>
        <align:entity1 rdf:resource="http://example.org/cmt#Paper"/>
        <align:entity2 rdf:resource="http://example.org/conference#Paper"/>
        <align:measure>1.0</align:measure>
        <align:relation>equivalent</align:relation>
    </align:cell>
    <align:cell>
        <align:entity1 rdf:resource="http://example.org/cmt#Author"/>
        <align:entity2 rdf:resource="http://example.org/conference#Author"/>
        <align:measure>1.0</align:measure>
        <align:relation>equivalent</align:relation>
    </align:cell>
    <align:cell>
        <align:entity1 rdf:resource="http://example.org/cmt#Review"/>
        <align:entity2 rdf:resource="http://example.org/conference#Review"/>
        <align:measure>1.0</align:measure>
        <align:relation>equivalent</align:relation>
    </align:cell>
    <align:cell>
        <align:entity1 rdf:resource="http://example.org/cmt#ProgrammeCommittee"/>
        <align:entity2 rdf:resource="http://example.org/conference#ProgrammeCommittee"/>
        <align:measure>1.0</align:measure>
        <align:relation>equivalent</align:relation>
    </align:cell>
</align:Alignment>
```

S2B- Ontology Alignment- Baseline prompt (metadata)

```
You are given two ontologies describing academic
conference organisation. The source ontology is cmt
and the target ontology is the conference ontology.
Both describe concepts related to academic events
including papers, authors, reviews, and programme
committees.

Generate an ontology alignment that identifies
equivalent concepts between the two ontologies.

Return the alignment in RDF/XML format using the
Alignment API format with align:Alignment structure.
Each correspondence must use align:Cell with
align:entity1, align:entity2, align:measure,
and align:relation.
The following lifecycle metadata is drawn from the Mapping
Metadata Vocabulary (MMV: http://w3id.org/mmv/0.1). Use it
to guide your alignment generation decisions.

[mmv:AnalysisActivity]
mmv:hasPurpose: Enable data integration between two
```

```
conference management systems by identifying semantically
equivalent concepts across their respective ontologies.
mmv:hasMappingDomain: Academic / Conference Organisation
mmv:hasRequirement: Each correspondence must reference
real URIs from the actual cmt and conference ontologies.
Each Cell must include entity1, entity2, measure,
and relation.
mmv:hasAssumption: Equivalence is based on semantic
meaning not just label similarity. Some concepts in
cmt may have no equivalent in the conference ontology.
mmv:hasDomainAssumption: The cmt ontology uses the
namespace http://cmt# for example http://cmt#Paper,
http://cmt#Author, http://cmt#Review,
http://cmt#Reviewer, http://cmt#ProgramCommittee.
The conference ontology uses the namespace
http://conference# for example http://conference#Paper,
http://conference#Author, http://conference#Review,
http://conference#Programme_committee.
mmv:hasTechnicalRequirement: Valid Alignment API
RDF/XML. Relation must use = not the word equivalent.
mmv:hasRiskOrIssue: Using fabricated URIs like
http://example.org/cmt# produces an alignment that
scores zero against the OAEI gold standard.
mmv:hasInputName: cmt.owl and conference.owl
mmv:hasInputFormat: OWL/RDF

[mmv:DesignActivity]
mmv:hasDesignDecision: Use the actual cmt and
conference ontology namespaces. Assign varied
confidence values reflecting semantic similarity
strength. Use = as the relation for equivalence.
mmv:hasJustification: The alignment will be evaluated
against the OAEI conference gold standard. Only
correspondences using real ontology URIs can score
against the gold standard.
mmv:hasPlannedQualityMetric: Precision and Recall
against OAEI conference gold standard. All URIs
must use real ontology namespaces. Confidence
values must vary between 0 and 1.

[mqv:DevelopmentActivity]
mqv:hasMappingType: Ontology Alignment
mqv:hasMappingFormat: Alignment API RDF/XML

After generating the alignment append a metadata
block using MMV:
```

S2B- Ontology Alignment- Baseline prompt result (metadata)

```
@prefix align: <http://knowledgeweb.semanticweb.org/heterogeneity/alignment#> .
@prefix cmt: <http://cmt#> .
@prefix conference: <http://conference#> .
```

```
@prefix mmv: <http://w3id.org/mmv/0.1#> .
@prefix mqv: <https://alex-randles.github.io/MQV/#> .

<http://example.org/alignment/oaei-conference>
    a mmv:MappingArtifact ;
    mmv:hasPurpose "Conference ontology alignment for academic data integration." ;
    mmv:hasMappingDomain "Academic / Conference Organisation" ;
    mmv:hasDesignDecision "Use real cmt and conference ontology URIs. Use = for equivalence.
Vary confidence values." ;
    mmv:hasJustification "Enables evaluation against OAEI conference gold standard." ;
    mqv:hasMappingType "Ontology Alignment" ;
    mqv:hasMappingFormat "Alignment API RDF/XML" .

align:Alignment a align:Alignment ;
    align:cell [
        align:entity1 cmt:Paper ;
        align:entity2 conference:Paper ;
        align:measure 0.9 ;
        align:relation "="
    ] ;
    align:cell [
        align:entity1 cmt:Author ;
        align:entity2 conference:Author ;
        align:measure 0.95 ;
        align:relation "="
    ] ;
    align:cell [
        align:entity1 cmt:Review ;
        align:entity2 conference:Review ;
        align:measure 0.85 ;
        align:relation "="
    ] ;
    align:cell [
        align:entity1 cmt:ProgramCommittee ;
        align:entity2 conference:Programme_committee ;
        align:measure 0.8 ;
        align:relation "="
    ] .
```

Interliking

S3A- Interlinking- Baseline prompt (no metadata)

```
1.  Generate a Interlinking mapping artefact for the following task.
1.  Task description:
2.  You are given two RDF datasets containing information
3.  about Irish places.
4.  Dataset A contains records from Ordnance Survey Ireland
5.  with properties: rdfs:label, dbo:location,
6.  geo:lat, geo:long.
7.  Dataset B contains Wikidata records with properties:
8.  rdfs:label, wdt:P17, geo:lat, geo:long.
9.  Generate a SPARQL query that links matching places
10. across the two datasets using owl:sameAs.
11. Return only the SPARQL query.
12. Return only the Interlinking  output with no extra explanation.
```

S3A- Interlinking- Baseline prompt result (no metadata)

```
1.  PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2.  PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
3.  PREFIX dbo: <http://dbpedia.org/ontology/>
4.  PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
5.  PREFIX wdt: <https://www.wikidata.org/prop/direct/>
6.  PREFIX owl: <http://www.w3.org/2002/07/owl#>
7.
8.  CONSTRUCT {
9.    ?placeA owl:sameAs ?placeB .
10. }
11. WHERE {
12.   ?placeA a dbo:Place ;
13.           rdfs:label ?labelA ;
14.           geo:lat ?latA ;
15.           geo:long ?longA .
16.
17.   ?placeB a wdt:P17 ;
18.           rdfs:label ?labelB ;
19.           geo:lat ?latB ;
20.           geo:long ?longB .
21.
22.   FILTER (str(?labelA) = str(?labelB) && ?latA = ?latB && ?longA = ?longB)
23. }
```

S3A- Interlinking- Baseline prompt (metadata)

```
1.  Generate a Interlinking mapping artefact for the following task.
2.
3.  Task description:
```

```
You are given two RDF datasets containing information
about Irish places.

Dataset A contains records from Ordnance Survey Ireland
with properties: rdfs:label, dbo:location,
geo:lat, geo:long.

Dataset B contains Wikidata records with properties:
rdfs:label, wdt:P17, geo:lat, geo:long.

Generate a SPARQL query that links matching places
across the two datasets using owl:sameAs.
Return only the SPARQL query.

Return only the Interlinking output with no extra explanation.

Lifecycle metadata:
The following lifecycle metadata is drawn from the Mapping
Metadata Vocabulary (MMV: http://w3id.org/mmv/0.1). Use it
to guide your interlinking query generation decisions.

[mmv:AnalysisActivity]
mmv:hasPurpose: Enrich the Virtual Record Treasury of
Ireland knowledge graph by linking OSi place records
to Wikidata entries to enable access to additional
contextual data about Irish historical places.
mmv:hasMappingDomain: Geospatial / Irish Historical Geography
mmv:hasRequirement: Query must use at least two
independent matching signals. Must produce owl:sameAs
triples. Must handle case-insensitive name matching.
mmv:hasAssumption: Name matching alone is insufficient
due to spelling variants and common place names.
County information provides supporting evidence.
Coordinates can disambiguate places with the same name.
mmv:hasDomainAssumption: Exact coordinate equality is
too strict for geospatial matching due to differences
in coordinate precision between datasets. A tolerance
of 0.05 decimal degrees should be applied using FILTER
with ABS(). Case-insensitive comparison is required
using lcase() for name matching.
mmv:hasTechnicalRequirement: Valid executable SPARQL
query. Must use FILTER for coordinate tolerance.
Must produce owl:sameAs assertions.
mmv:hasRiskOrIssue: Exact string matching on labels
produces false negatives for spelling variants.
Exact coordinate equality produces false negatives
due to coordinate precision differences between
OSi and Wikidata datasets.
mmv:hasInputFormat: RDF/Turtle

[mmv:DesignActivity]
mmv:hasDesignDecision: Use multi-signal matching
```

```
combining case-insensitive name similarity and
coordinate proximity within 0.05 decimal degree
tolerance. Use lcase() for name comparison.
Use ABS() with FILTER for coordinate tolerance.
mmv:hasJustification: Single-signal matching on
name alone produces false positives for common
Irish place names. The combination of name and
coordinate evidence is standard practice in
geospatial link discovery as used in LIMES and Silk.
mmv:hasPlannedQualityMetric: Query uses minimum
two matching signals. Coordinate tolerance
explicitly defined as 0.05. Query executes
without syntax errors and returns results
against sample data.

[mqv:DevelopmentActivity]
mqv:hasMappingType: Interlinking
mqv:hasMappingFormat: SPARQL CONSTRUCT query

After generating the SPARQL query append a
metadata block using MMV:
Return only the Interlinking output with no extra explanation.
```

S3A- Interlinking- Baseline prompt result (metadata)

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX wdt: <https://www.wikidata.org/prop/direct/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

CONSTRUCT {
  ?placeA owl:sameAs ?placeB .
}
WHERE {
  ?placeA a dbo:Place ;
          rdfs:label ?labelA ;
          geo:lat ?latA ;
          geo:long ?longA .

  ?placeB a wdt:P17 ;
          rdfs:label ?labelB ;
          geo:lat ?latB ;
          geo:long ?longB .

  FILTER (lcase(str(?labelA)) = lcase(str(?labelB)) &&
          ABS(?latA - ?latB) <= 0.05 &&
          ABS(?longA - ?longB) <= 0.05)
}

@prefix mmv: <http://w3id.org/mmv/0.1#> .
```

```
@prefix mqv: <https://alex-randles.github.io/MQV/#> .

<http://example.org/mapping/vrti-interlink>
    a mmv:MappingArtifact ;
    mmv:hasPurpose "Link OSi Irish places to Wikidata for VRTI knowledge graph enrichment." ;
    mmv:hasMappingDomain "Geospatial / Irish Historical Geography" ;
    mmv:hasDesignDecision "Multi-signal matching: lcase() name match + coordinate tolerance 0.05." ;
    mmv:hasJustification "Prevents false positives from name-only matching on common Irish place names." ;
    mqv:hasMappingType "Interlinking" ;
    mqv:hasMappingFormat "SPARQL CONSTRUCT query" .
```

S3B- Interlinking- Baseline prompt (no metadata)

```
Generate a Interlinking mapping artefact for the following task.

Task description:
You are given two RDF datasets containing geographic
place information.

Dataset A is from DBpedia with properties:
rdfs:label, dbo:country, geo:lat, geo:long,
rdf:type dbo:Place.

Dataset B is from GeoNames with properties:
gn:name, gn:countryCode, geo:lat, geo:long,
rdf:type gn:Feature.

Generate a SPARQL query that identifies matching
places across both datasets and generates
owl:sameAs links.
Return only the SPARQL query.

Return only the Interlinking output with no extra explanation.
```

S3B- Interlinking- Baseline prompt result (no metadata)

```
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX gn: <http://www.geonames.org/ontology#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

CONSTRUCT {
  ?placeA owl:sameAs ?placeB .
}
WHERE {
  ?placeA a dbo:Place ;
          rdfs:label ?labelA ;
```

```
            dbo:country ?countryA ;
            geo:lat ?latA ;
            geo:long ?longA .

   ?placeB a gn:Feature ;
            gn:name ?labelB ;
            gn:countryCode ?countryB ;
            geo:lat ?latB ;
            geo:long ?longB .

   FILTER (str(?labelA) = str(?labelB) &&
           ?countryA = ?countryB &&
           ?latA = ?latB &&
           ?longA = ?longB)
}
```

S3B- Interlinking- Baseline prompt (metadata)

```
Generate a Interlinking mapping artefact for the following task.

Task description:
You are given two RDF datasets containing geographic
place information.

Dataset A is from DBpedia with properties:
rdfs:label, dbo:country, geo:lat, geo:long,
rdf:type dbo:Place.

Dataset B is from GeoNames with properties:
gn:name, gn:countryCode, geo:lat, geo:long,
rdf:type gn:Feature.

Generate a SPARQL query that identifies matching
places across both datasets and generates
owl:sameAs links.
Return only the SPARQL query.

Lifecycle metadata:
The following lifecycle metadata is drawn from the Mapping
Metadata Vocabulary (MMV: http://w3id.org/mmv/0.1). Use it
to guide your interlinking query generation decisions.

[mmv:AnalysisActivity]
mmv:hasPurpose: Enrich DBpedia place entities with
GeoNames identifiers to improve geographic coverage
and enable cross-dataset queries over complementary
place descriptions.
mmv:hasMappingDomain: Geospatial / Linked Open Data
mmv:hasRequirement: Query must use at least two
independent matching signals. Must produce owl:sameAs
triples. Must handle case-insensitive name matching.
mmv:hasAssumption: Place name matching must be
```

```
case-insensitive using lcase(). DBpedia country
values are URIs such as dbr:Ireland. GeoNames
countryCode values are ISO 3166-1 alpha-2 strings
such as IE. These cannot be compared directly and
should not be used as a matching signal.
mmv:hasDomainAssumption: Coordinate proximity within
0.05 decimal degrees indicates a likely match.
Exact coordinate equality is too strict due to
differences in coordinate precision between
DBpedia and GeoNames. ABS() with FILTER must be
used for coordinate matching.
mmv:hasTechnicalRequirement: Valid executable SPARQL
query. Must use FILTER with lcase() for name matching.
Must use ABS() for coordinate tolerance.
Must produce owl:sameAs triples.
mmv:hasRiskOrIssue: Comparing DBpedia country URIs
directly to GeoNames ISO country code strings will
always fail and produce zero results. Exact coordinate
equality produces false negatives due to precision
differences between datasets.
mmv:hasInputFormat: RDF

[mmv:DesignActivity]
mmv:hasDesignDecision: Use two matching signals ,Äî
case-insensitive name matching using lcase() and
coordinate proximity using ABS() with tolerance
0.05. Do not use country as a matching signal
because DBpedia and GeoNames represent country
in incompatible formats.
mmv:hasJustification: Country matching between
DBpedia URIs and GeoNames ISO codes will always
fail. Name and coordinate signals are sufficient
for reliable geospatial interlinking and are
consistent with standard practice in LIMES
and Silk link discovery tools.
mmv:hasPlannedQualityMetric: Query uses name
and coordinate signals only. Coordinate tolerance
0.05 explicitly stated. Query returns results
when tested against sample data. Does not
attempt country URI to string comparison.

[mqv:DevelopmentActivity]
mqv:hasMappingType: Interlinking
mqv:hasMappingFormat: SPARQL CONSTRUCT query

After generating the SPARQL query append a
metadata block using MMV:
Return only the Interlinking output with no extra explanation.
```

S3B- Interlinking- Baseline prompt result (metadata)

```
PREFIX dbo: <http://dbpedia.org/ontology/>
```

```
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX gn: <http://www.geonames.org/ontology#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

CONSTRUCT {
  ?placeA owl:sameAs ?placeB .
}
WHERE {
  ?placeA a dbo:Place ;
           rdfs:label ?labelA ;
           geo:lat ?latA ;
           geo:long ?longA .

  ?placeB a gn:Feature ;
           gn:name ?labelB ;
           geo:lat ?latB ;
           geo:long ?longB .

  FILTER (lcase(?labelA) = lcase(?labelB))
  FILTER (ABS(?latA - ?latB) < 0.05)
  FILTER (ABS(?longA - ?longB) < 0.05)
}

@prefix mmv: <http://w3id.org/mmv/0.1#> .
@prefix mqv: <https://alex-randles.github.io/MQV/#> .

<http://example.org/mapping/dbpedia-geonames>
    a mmv:MappingArtifact ;
    mmv:hasPurpose "Link DBpedia places to GeoNames for geographic coverage enrichment." ;
    mmv:hasMappingDomain "Geospatial / Linked Open Data" ;
    mmv:hasDesignDecision "Name lcase() match + coordinate ABS() tolerance 0.05. Country
signal excluded due to incompatible formats." ;
    mmv:hasJustification "DBpedia country URIs and GeoNames ISO codes cannot be compared
directly." ;
    mqv:hasMappingType "Interlinking" ;
    mqv:hasMappingFormat "SPARQL CONSTRUCT query" .
```

Towards a Unified Metadata Model for Semantic and Data Mappings

Sarah Alzahrani[1,2], Declan Osullivan[3]

[1]School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland
[2]College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia
[3]ADAPT Centre for Digital Content, School of Computer Science and Statistics, Trinity College Dublin, Dublin , Ireland

Abstract

Declarative data mappings are a valuable method for specifying how two datasets relate to one another. Creating these mappings is not a simple process, since they are typically intended for a particular data integration purpose and require expertise, resources, and time to accurately compose. However, there are still many challenges facing researchers regarding finding, accessing, understanding, and reusing these mappings, including the absence of detailed documentation and metadata related to their development lifecycle. In this paper, we argue that a unified metadata model that captures every aspect of the mapping development lifecycle is the most effective means of overcoming the current limitations

Keywords

Mappings, Alignments, Metadata Model, Knowledge Representation

1. Introduction

Within the linked data and semantic web domain mapping processes can be classified into three categories, namely interlinking mapping, uplift/downlift mapping, and ontologies mapping (sometimes called ontology alignment). The process of creating these mappings can be a complex and time-consuming task [1]. Thus, the reuse of previously developed mappings is considered necessary. It is our intuition that annotating mappings with a unified machine-processable metadata model describing the development lifecycle may enhance the discovery, reuse, and quality assessment of such mappings. This paper presents an overview of our proposed mapping lifecycle and a description of our initial metadata model.

2. Proposed Mapping lifecycle

A mapping lifecycle breaks down the mapping process into stages. The proposed lifecycle in this research has been designed such that it will be easy to follow by engineers who may not have a knowledge engineering background but who may have a more general software engineering background. The lifecycle is composed of six phases: Analysis, Mapping Design, Mapping

OM-2022: Proceedings of the 17th International Workshop on Ontology Matching, October 2022, Hangzhou, China (Virtual)
salzahra@tcd.ie (S. Alzahrani); declan.osullivan@tcd.ie (D. Osullivan)
0000-0003-2169-7006 (S. Alzahrani); 0000-0003-1090-3548 (D. Osullivan)
© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3324/om2022_poster2.pdf

Table 1
Proposed metadata fields

Phases	Proposed metadata fields
Analysis	Metadata about stakeholders (URI, Name, Background, Role, organization), Metadata about the purpose of the mapping (Requirements, Type of mapping, Mapping domain, Domain assumptions, technical requirement, Risks, or issues), and Metadata to describe inputs that will be mapped (URI, Name, Source, Type, Creator, Format)
Design	(Final design decisions,Design decision justification,Quality metrics to consider during the development)
Development	(URI, Name, Start/End date of the development, Tools, Mapping method, Mapping algorithm, Format)
Testing	(URI, Name, Testing type, Date/Time, Testing result)
Maintenance	(Publisher name, Publisher source, Version number, Version date/time)

Development, Mapping Testing, and Map-ping Maintenance. We argue that the lifecycle phases are appropriate irrespective of which of the three types of declarative mappings (ontology, uplift/downlift, interlinking) is being handled, but naturally, some parts are more relevant/less relevant depending on the type of mapping.

3. Initial Metadata Model

Several studies in the literature have introduced different metadata models and ways to enrich mappings with metadata [2][3]. Generally, previous studies have developed metadata models for specific types of mappings, and a few have considered comprehensive metadata fields reflecting the stages of mapping development. In previous research such as [1] and [4], we have considered the availability of comprehensive metadata fields reflecting the different stages of mapping development for ontologies mappings. In this study, as an initial step toward supporting mapping-related activities, we have developed a unified metadata model that captures the lifecycle of mapping development. A summary of the metadata fields for each phase in the proposed lifecycle is shown in the table below.

4. Future work

The next step in our research is seeking feedback from the community through a survey. The feedback gathering is still ongoing, and 18 responses have been received to date. Your participation is appreciated at https://rb.gy/kfpcrk. Next, we'll examine how the metadata will be represented and attached to mappings. Initially, we will investigate RDF* and Named graphs as possible ways to represent our metadata model. Following on, an annotation tool will be developed, and its usability and accuracy will be evaluated by engineers. Lastly, we believe this work will contribute to building a collaborative mapping platform where semantic web and linked data community members can share, update, and evaluate mappings.

5. Acknowledgments

The first author acknowledges financial support from Imam Mohammad Ibn Saud Islamic University and the Saudi government represented by the Royal Embassy of Saudi Arabia Cultural Bureau in Dublin. The second author is partially supported by the SFI ADAPT Research Centre (grant number 13/RC/2106_P2).

References

[1] C. Debruyne, B. Walshe, and D. O'Sullivan, "Towards a Project Centric Metadata Model and Lifecycle for Ontology Mapping Governance," in Proceedings of the 17th International Conference on Information Integration and Web-based Applications Services, New York, NY, USA, 2015, vol. 1–10, p. Article 50. doi: 10.1145/2837185.2837201.

[2] A. Dimou, T. De Nies, R. Verborgh, E. Mannens, P. Mechant, and R. Van de Walle,"Automated metadata generation for Linked Data generation and publishing workflows" in LDOW2016, 2016, pp.1–10.

[3] N. Matentzoglu et al., "A Simple Standard for Sharing Ontological Mappings (SSSOM)," Data-base, vol. 2022, p. baac035, Jan. 2022, doi: 10.1093/database/baac035.

[4] H. Thomas, R. Brennan, and D. O'Sullivan, "Using the OM2R Meta-Data Model for Ontology Mapping Reuse for the Ontology Alignment Challenge: A Case Study," in Proceedings of the 7th Interna-tional Conference on Ontology Matching - Volume 946, Aachen, DEU, 2012, pp. 61–72.

What metadata is needed for semantic and data mappings?

Sarah Alzahrani
Computer Science and Statistics
Trinity College Dublin
Dublin, Ireland
salzahra@tcd.ie

Declan O'Sullivan
ADAPT Centre, Computer Science and Statistics
Trinity College Dublin
Dublin, Ireland
declan.osullivan@tcd.ie

2023 IEEE 17th International Conference on Semantic Computing (ICSC) | 978-1-6654-8263-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSC56153.2023.00064

***Abstract*— A huge amount of data is produced each day through social networks like Facebook, Twitter, etc. Through semantic web technologies, data produced from these social networks can be shared, integrated, and used to gain knowledge. As social networks grow more heterogeneous, declarative data mapping can help address this issue. The use of declarative data mappings is a valuable method of describing the relationship between two different datasets. Theses mappings are created within the community, but without standard metadata such that quality mappings cannot be shared, discovered, and reused properly. The ultimate purpose of our research is to identify a set of metadata fields/model that can be used to annotate declarative semantic and data mappings to facilitate their discovery, reusability, and quality assessment. As an early step in our research, 18 participants completed an online questionnaire between June and September 2022. The questionnaire targeted participants with varying levels of experience with semantic and data mappings. The survey results indicate that the model we have proposed for mapping metadata has received a positive response from the expert and intermediate-level participants. Also, expert participants have suggested adding extra metadata fields which are being considered as the second version of the model that is being developed.**

Keywords*— *Mapping, Alignment, Metadata Model, Knowledge Representation, Annotation.

I. INTRODUCTION

The World Wide Web (WWW) is widely used as a means of exchanging information. Interoperability between different information systems on the web is limited due to semantic heterogeneity and heterogeneous data. This issue can be observed when working with different datasets and ontologies, as well as when publishing data to the web. As a solution to this issue, declarative mappings have been suggested as a useful method of specifying how two semantics or data models relate to one another [1]. Within linked data and semantic web domains, declarative mapping processes can be classified into three categories, namely ontology mapping, interlinking, and uplift/downlift. The uplift mapping process involves converting different data formats to the Resource Description Framework (RDF), while the downlift mapping process refers to the reverse [2]. Interlinking involves linking named resources (identified by URIs) in order to indicate how they are related, including if they describe the same concept [3]. Ontology mapping enables the discovery of semantically similar correspondences between different ontologies [1]. These three mapping processes can be found in social networks, for example, when linking social network data sets, mapping similar ontologies, and elevating social network data into RDF. All three mapping processes result in mappings expressed in a declarative manner, although the representation of the declarative mappings may differ. The process of creating these mappings is not trivial since they are intended for a specific purpose and demand expertise, resources, and time [4]. To facilitate their development and reduce the associated cost, the re-use of previously developed declarative mappings is considered necessary. It is argued that identifying appropriate mappings for reuse can be accomplished by examining the metadata associated with these mappings.

Metadata is often referred to as data about data. A clear acknowledgment of the importance of metadata to describe a mapping is found in [5], where Shvaiko emphasizes the need to carefully consider alignment metadata and annotations to provide standard ways to communicate and retrieve alignments. There have been several studies in the literature that have introduced different metadata models and ways to enrich mappings with metadata [6][7][8][4]. Most of these studies provide metadata that focuses only on the final generated mapping. However, a few studies, including [4] and [8], have taken into account the entire lifecycle involved in developing a mapping, not just its final outcome. Furthermore, previous studies have developed metadata models for specific types of mappings, and to the best of the authors' knowledge no previous work have consider a unified metadata model for all type of mappings. Moreover, most existing declarative mappings tend to have no metadata or if they do, have minimal metadata represented as plain text (making them difficult to query and process). These limitations make it challenging for academia and industry to discover and reuse mappings effectively and efficiently.

978-1-6654-8263-9/23/$31.00 ©2023 IEEE
DOI 10.1109/ICSC56153.2023.00064

Authorized licensed use limited to: TRINITY COLLEGE LIBRARY DUBLIN. Downloaded on April 21,2026 at 04:16:20 UTC from IEEE Xplore. Restrictions apply.

The ultimate purpose of our research is to identify a set of metadata fields/model that can be used to annotate declarative semantic and data mappings to facilitate their discovery, reusability, and quality assessment. This paper describes the results of an initial survey we have undertaken to evaluate our first proposed metadata model, as described in [9]. In section II an overview of the survey methodology is presented, as well as a brief overview of our proposed metadata model that participants were asked to provide feedback upon. The survey results are presented in section III, followed by section IV, which discusses the results and highlights the main findings. The final section summarizes the results and discusses the future directions of this study.

II. METHOD

In our research, we are attempting to create an approach that will enable the annotation of mappings with a unified machine-readable metadata model that captures the development lifecycle of declarative mappings, irrespective of type. The purpose of the survey described in this paper was to get feedback on our first proposed model [9] with semantic web and linked data community members by examining how relevant the proposed set of metadata fields is for two mapping-related tasks, namely i) finding mappings that can be reused, and ii) evaluating mapping quality. These metadata fields (see Table 1) were identified to be associated with different phases of a proposed declarative mapping lifecycle. A mapping lifecycle breaks down the mapping process into stages that describe the construction of the mappings [9]. six phases: Analysis, Mapping Design, Mapping Development, Mapping Testing, and Mapping Maintenance.

A. Participants

Participation invitations were sent to researchers in the declarative mapping area and through community mailing lists, such as semantic-web@w3.org, the former semantic web interest group. A total of 18 participants agreed to participate in the study so far in the period July 2022 to September 2022. As one of the objectives of this study was to examine whether prior experience influences a person's assessment of the relevancy of metadata fields to mapping-related tasks, participants were asked to describe their level of experience regarding declarative semantic and data mappings. The first question of the survey asked how would you describe your current level of experience in data and semantic mapping. The participants had a choice between no experience, beginners, intermediates, and experts. **Nine described themselves as experts** in semantic and data mapping, **six as intermediates**, and **three as beginners**. However please note that when we analyzed the ratings provided by the 'beginners' cohort, no obvious tendency or common points of agreement could be detected in their ratings of the fields for either task, so we do not present these results as tables in section III, but rather just provide commentary on their results in section IV.

B. Procedure

The survey feedback was collected using a google form that was sent to participants via mailing lists. At the beginning of the survey as explained previously participants were asked to describe their current level of data and semantic mapping experience. Then they were asked to rate the relevance of various metadata fields to the two mapping-related tasks which were presented as: finding mappings that can be reused, and evaluating quality of a declarative mapping. These tasks were selected based on their importance in a variety of other mapping-related activities, such as supporting the reuse of high-quality mappings, creating new mappings, and maintaining them.

Each task was represented in separate section of the form. The first section of the questionnaire asked participants to rate the relevance of metadata fields for the mapping discovery process, based on their experience. Our provided definition stated that *The metadata fields relevant to the discovery process are expected to facilitate the discovery process when a query is made.* Participants were asked to rate (in their opinion) how relevant the proposed metadata fields are to the mapping discovery process on a scale of 1 to 5. The proposed metadata fields in this section are grouped into 7 categories, each of which corresponds to a phase of the mapping process (see Table 1). The second section of the questionnaire asked participants to rate the relevance of metadata fields for the quality assessment process, based on their experience. Our provided definition stated that *the metadata fields relevant to the quality assessment process are expected to facilitate the quality assessment process when a query is made*. Participants were asked to rate (in their opinion) how relevant the proposed metadata fields are to the mapping quality assessment process on a scale of 1 to 5. The proposed metadata fields in this section are grouped into 7 parts as well , each of which corresponds to a phase of the mapping process. Each section concluded with a question asking participants whether they could suggest any other metadata fields to improve the task.

TABLE 1. PROPOSED METADATA FIELDS

MAPPING LIFECYCLE PHASES	PROPOSED METADATA FIELDS
Analysis	Metadata about stakeholders (URI, Name, Background, Role, organization)
	Metadata about the purpose of the mapping (Requirements, Type of mapping, Mapping domain, Domain assumptions, technical requirement, Risks or issues)
	Metadata to describe inputs that will be mapped (URI, Name, Source, Type, Creator, Format).
Design	Final design decisions, Design decision justification, Quality metrics to consider during the development
Development	(URI, Name, Start/End date of the development, Tools, Mapping method, Mapping algorithm, Format)
Testing	(URI, Name, Testing type, Date/Time, Testing result)
Maintenance	(Publisher name, Publisher source, Version number, Version date/time).

III. RESULTS

The following is an analysis of the questionnaire responses received so far.

Authorized licensed use limited to: TRINITY COLLEGE LIBRARY DUBLIN. Downloaded on April 21,2026 at 04:16:20 UTC from IEEE Xplore. Restrictions apply.

A. Task: support finding mapping that can be reused

The following tables summarize the number of fields that were rated as relevant or not for the first task by each category of participants. As mentioned previously, participants were asked to rate (in their opinion) the relevance of proposed metadata fields to the mapping discovery process.

a) Ratings of proposed Metadata fields by Experts

In table 2 and 3 below, the number in the cell indicates the number of people who provided the rating shown in the column header for the metadata field whose name is given at the start of the row in question. So, for example, Table 2 shows that four experts rated the proposed field *Name* for describing stakeholders as relevant (taking together ratings 1 and 2 together) to the task. Further, it was rated neutral (rating 3) by two experts, less relevant by one expert, and irrelevant by two experts. As shown in Table 2, six fields have been proposed to describe stakeholders, namely URI, name, background, role, and organization.

TABLE 2. RESULTS FOR THE MAPPING DISCOVERY TASK -EXPERTS

Mapping discovery task - Experts						
	Proposed fields	**1 Highly relevant**	**2**	**3**	**4**	**5 irrelevant**
Metadata fields describing stakeholders	Name	2	2	2	1	2
	URI	3	1	1	3	1
	Role	2	3	3	1	0
	Background	1	1	3	4	0
	organization	2	1	5	1	0
Metadata fields describing the mapping project's purpose	Requirement	**5**	3	0	0	0
	Type of mapping	6	3	0	0	0
	Domain	4	3	2	0	0
	assumptions	2	4	3	0	0
	Technical Requirement	2	3	4	0	0
	Risks/issues	1	4	2	2	0
Metadata fields describing the input to be mapped	Name	1	4	2	1	1
	URI	**5**	1	1	1	1
	Source	**5**	1	2	1	0
	Type	**7**	0	2	0	0
	Creator	1	3	1	3	1
	Format	4	1	3	0	1
Metadata fields describing the mapping design process	Design decisions	4	2	2	0	1
	Justifications	3	3	2	1	0
	Quality Metrics	4	3	2	0	0
Metadata fields describing the mapping development process	Mapping Name	2	3	2	0	2
	Mapping URI	**6**	0	2	0	1
	Start date and time	1	3	2	1	2
	End date and time	1	4	2	1	1
	Used tools	3	1	4	0	1
	Mapping method	**5**	3	1	0	0
	Mapping algorithm	3	3	2	1	0
	Mapping format	**6**	2	1	0	0
Metadata fields describing the testing process of a mapping	Testing Identifier	3	0	3	1	2
	Testing Name	1	2	4	2	0
	Testing type	3	3	3	0	0
	Testing data and time	1	2	3	2	1
	Testing result	4	3	2	0	0
Metadata fields describing the the management and maintenance of mappings	Publisher name	4	3	1	0	1
	Publisher source	**5**	0	3	0	1
	Version number	**5**	2	0	1	1
	Version date/time	**5**	3	0	0	1

b) Ratings of proposed Metadata fields by Intermediates

TABLE 3. RESULTS FOR THE MAPPING DISCOVERY -INTERMEDIATE

Mapping discovery task – intermediates						
	Proposed fields	**1 Highly relevant**	**2**	**3**	**4**	**5 irrelevant**
Metadata fields describing stakeholders	Name	2	**3**	0	1	0
	URI	2	2	1	1	0
	Role	1	**3**	0	2	0
	Background	1	**3**	0	2	0
	organization	1	2	2	1	0
Metadata fields describing the mapping project's purpose	Requirement	**4**	1	0	1	0
	Type of mapping	**4**	1	1	0	0
	Domain	**5**	1	0	0	0
	assumptions	2	**3**	0	1	0
	Technical Requirement	2	2	0	2	0
	Risks/issues	2	2	0	2	0
Metadata fields describing the input to be mapped	Name	**3**	**3**	0	0	0
	URI	**4**	2	0	0	0
	Source	**4**	1	1	0	0
	Type	**3**	1	2	0	0
	Creator	1	2	2	1	0
	Format	1	2	**3**	0	0
Metadata fields describing the mapping design process	Design decisions	0	**3**	**3**	0	0
	Justifications	0	**3**	**3**	0	0
	Quality Metrics	0	**4**	2	0	0
Metadata fields describing the mapping development process	Mapping Name	2	**3**	0	0	1
	Mapping URI	2	**3**	0	0	1
	Start date and time	1	0	2	2	1
	End date and time	2	0	1	2	1
	Used tools	1	**3**	0	2	0
	Mapping method	1	**4**	1	0	0
	Mapping algorithm	1	2	**3**	0	0
	Mapping format	1	1	**4**	0	0
Metadata fields describing the testing process of a mapping	Testing Identifier	2	2	0	2	0
	Testing Name	1	1	2	2	0
	Testing type	**3**	2	0	1	0
	Testing data and time	1	1	1	2	1
	Testing result	3	2	1	0	0
Metadata fields describing the the management and maintenance of mappings	Publisher name	2	2	2	0	0
	Publisher source	**3**	1	2	0	0
	Version number	1	**3**	2	0	0
	Version date/time	**3**	0	**3**	0	0

c) other metadata fields recommened by participants to improve the mapping discovery process

There have been six recommendations from experts for adding more metadata fields to the model. Two of the participants suggested reviewing the previously proposed Simple Standard for Sharing Ontology Mappings (SSSOM), which is a simple metadata standard for describing semantic mappings [6]. Several participants have suggested additional fields including publisher URI, mapping constructs, number of projects, data amount, and the change/issue management process.

B. Task: support evaluation of the quality of mapping

The following tables summarize the number of fields that were rated as relevant or not for the mapping quality assessment task by each category of participants.

Authorized licensed use limited to: TRINITY COLLEGE LIBRARY DUBLIN. Downloaded on April 21,2026 at 04:16:20 UTC from IEEE Xplore. Restrictions apply.

a) Ratings of proposed Metadata fields by Experts

In table 4 and 5 below, the number in the cell indicates the number of people who provided the rating shown in the column header for the metadata field whose name is given at the start of the row in question.

TABLE 4. RESULTS FOR THE MAPPING QUALITY ASSESSMENT TASK - EXPERTS

Mapping quality assessment task - Experts						
	Proposed fields	**1 Highly relevant**	**2**	**3**	**4**	**5 irrelevant**
Metadata fields describing stakeholders	Name	1	3	1	2	1
	URI	2	1	2	2	1
	Role	2	2	3	1	0
	Background	1	1	4	2	0
	organization	2	3	2	1	0
Metadata fields describing the mapping project's purpose	Requirement	**7**	1	0	0	0
	Type of mapping	**6**	1	0	1	0
	Domain	4	2	2	0	0
	assumptions	3	**5**	0	0	0
	Technical Requirement	3	3	1	1	0
	Risks/issues	1	4	1	2	0
Metadata fields describing the input to be mapped	Name	0	3	2	1	2
	URI	2	2	1	2	1
	Source	4	1	1	2	0
	Type	5	1	0	2	0
	Creator	1	2	0	4	1
	Format	3	1	1	2	1
Metadata fields describing the mapping design process	Design decisions	2	**5**	0	0	1
	Justifications	3	3	2	0	0
	Quality Metrics	4	4	0	0	0
Metadata fields describing the mapping development process	Mapping Name	0	3	1	2	2
	Mapping URI	4	1	1	1	1
	Start date and time	1	2	1	2	2
	End date and time	1	3	1	2	1
	Used tools	4	0	2	1	1
	Mapping method	4	4	0	0	0
	Mapping algorithm	3	3	2	0	0
	Mapping format	2	2	1	1	2
Metadata fields describing the testing process of a mapping	Testing Identifier	2	1	3	0	2
	Testing Name	0	3	3	0	2
	Testing type	**5**	1	2	0	0
	Testing data and time	1	4	1	1	2
	Testing result	**5**	2	1	0	0
Metadata fields describing the the management and maintenance of mappings	Publisher name	3	1	1	1	2
	Publisher source	2	0	4	1	1
	Version number	3	2	0	2	1
	Version date/time	3	3	0	0	2

b) Ratings of proposed Metadata fields by Intermediates

TABLE 5. RESULTS FOR QUALITY ASSESSMENT TASK INTERMEDIATE

Mapping quality assessment task – intermediates						
	Proposed fields	**1 Highly relevant**	**2**	**3**	**4**	**5 irrelevant**
Metadata fields describing stakeholders	Name	1	**3**	1	1	0
	URI	0	**3**	2	1	0
	Role	1	1	2	2	0
	Background	1	**4**	1	0	0
	organization	1	2	0	**3**	0
Metadata fields describing the mapping project's purpose	Requirement	2	**3**	0	1	0
	Type of mapping	2	**3**	0	1	0
	Domain	**3**	1	0	1	1
	assumptions	**3**	0	0	0	0
	Technical Requirement	2	**3**	0	1	0
	Risks/issues	2	2	2	0	0
Metadata fields describing the input to be mapped	Name	2	**4**	0	0	0
	URI	**3**	**3**	0	0	0
	Source	**5**	0	0	1	0
	Type	**3**	2	0	1	0
	Creator	2	1	2	1	0
	Format	1	**4**	0	1	0
Metadata fields describing the mapping design process	Design decisions	**4**	1	1	0	0
	Justifications	**4**	1	1	0	0
	Quality Metrics	**4**	1	1	0	0
Metadata fields describing the mapping development process	Mapping Name	2	2	2	0	0
	Mapping URI	2	2	2	0	0
	Start date and time	1	2	0	2	1
	End date and time	1	2	0	2	1
	Used tools	2	1	2	1	0
	Mapping method	**5**	0	1	0	0
	Mapping algorithm	**3**	2	1	0	0
	Mapping format	1	2	2	1	0
Metadata fields describing the testing process of a mapping	Testing Identifier	**4**	0	2	0	0
	Testing Name	**3**	0	**3**	0	0
	Testing type	**5**	0	1	0	0
	Testing data and time	**4**	0	1	1	0
	Testing result	**5**	1	0	0	0
Metadata fields describing the the management and maintenance of mappings	Publisher name	2	**3**	1	0	0
	Publisher source	3	1	0	2	0
	Version number	1	4	0	1	0
	Version date/time	3	1	0	2	0

c) other metadata fields recommended by participants to improve the qualirt assessment process

One of the expert participants proposed extra fields, including mapping cardinality, mapping justification, and mapping confidence from the SSOM model [6]

IV. DISCUSSION

Our research is focused on identifying a set of metadata fields for annotation of declarative semantic and data mappings to facilitate their discovery, reusability, and quality assessment. We undertook a feedback survey based on our initial model with members of the linked data and semantic web communities.

In the first task, which is to find a mapping that can be reused, **experts** (nine participants) have considered **the majority of the fields relevant** to the task but with three fields considered less relevant by four experts, namely *stakeholder URI*, *stakeholder background* and *input creator*. A total of six **intermediate-level** participants rated **all the fields relevant** to the task but with three fields considered less relevant, which are development of *mapping starting date*, *ending date*, and *testing date/time*. The three participants who were **beginners** rated **thirteen fields irrelevant**, including *stakeholder URIs, stakeholder backgrounds, stakeholder organizations, risks identified before development, input creators, input formats, development starts and ends date, tools used during development, mapping formats, testing ids, testing names, testing types, testing data and times, and testing results*.

Authorized licensed use limited to: TRINITY COLLEGE LIBRARY DUBLIN. Downloaded on April 21,2026 at 04:16:20 UTC from IEEE Xplore. Restrictions apply.

On the second task, assessing the quality of mapping, **experts** (eight participants) considered **all fields are relevant** but three less relevant, which *are input creator, mapping human readable name, and development of mapping starting date*. **Intermediate** participants, representing six individuals, rated **all fields relevant** to the task but three fields less relevant: *stakeholder organization, development of the mapping starting and ending date*. The **beginners**, which were two participants, rated **14 fields as irrelevant**, including *stakeholder's human readable name, stakeholder's URI, stakeholder's role, stakeholder backgrounds, stakeholder organizations, domain assumptions, risks identified before development, input creators, input formats, beginning and ending dates of development, tools used to develop, mapping formats, testing data and times, and publisher name*

From this analysis it is evident that there was an influence of a participant's level of expertise on his/her assessment of how relevant the fields are despite the small number of participants. A good example of this influence is when experts and intermediate-level participants, which constitute most of the group, rated only three fields as irrelevant for each task, while beginners rated about thirteen fields as irrelevant. Moreover, only experts have suggested extra metadata fields to the model, out of all the other participants.

To conclude, the main finding of this survey is that the model we developed received a reasonably positive response from the expert and intermediate-level participants, and they rated most of the proposed metadata fields as relevant to the tasks identified and suggested additional useful metadata fields. However, the different categories of participants did not agree on those fields that were deemed less relevant. To put it differently, there were no common fields among the fields that the experts and intermediate-level participants considered irrelevant, at this stage of development of the metadata model. In regard to the extra metadata fields proposed by the experts for both tasks, two experts have suggested considering the SSOM model [6] However the SSOM model focuses mainly on the actual mapping, rather than the mapping development lifecycle as a whole. Regardless, what has been suggested in their model can be reused and added to the group of metadata fields we propose to describe the mapping development process. As well as this, other experts have suggested adding the *publisher identifier, mapping construct, amount of data being mapped, and changes during the maintenance process*, which will all be considered when the second version of the model is developed.

V. CONCLUSION

This paper presents the results of a questionnaire survey on metadata model for declarative semantic and data mappings. It was conducted to verify a set of predefined metadata proposed by the authors as part of an attempt to develop a machine-readable metadata model that captures the entire mapping development lifecycle so that annotations of mappings can be generated and queried. In this study, 18 participants completed an online questionnaire between June and September 2022. The feedback gathering is ongoing. Your participation is appreciated at https://rb.gy/kfpcrk. The questionnaire targeted participants with varying levels of experience with semantic and data mappings from the linked data and semantic web community members. The initial survey results illustrate that the model we have developed has received a positive response from the expert and intermediate level participants, and they rated most of the proposed metadata fields as relevant to the tasks identified. In addition, expert participants have suggested adding extra metadata fields which will all be considered when the second version of the model is developed.

Next steps we will look at how the metadata will be represented and associated with declarative mappings. As a starting point, we will investigate RDF* and Named graphs as possible ways to represent our metadata model. In the following steps, an annotation tool will be developed, and its usability and accuracy will be evaluated by engineers. Finally, we hope to develop a collaborative mapping platform that allows semantic web and linked data experts to share, update, and evaluate mappings.

ACKNOWLEDGMENT

The first author acknowledges financial support from Al Imam Mohammad Ibn Saud Islamic University and the Saudi government - represented by the Royal Embassy of Saudi Arabia - Cultural Bureau in Dublin. The second author is partially supported by the SFI ADAPT Research Centre (grant number 13/RC/2106_P2).

REFERENCES

[1] M. Mao, "Ontology Mapping: An Information Retrieval and Interactive Activation Network Based Approach," in *The Semantic Web*, Berlin, Heidelberg, 2007, pp. 931–935.

[2] A. Junior, C. Debruyne, R. Brennan, and D. O'Sullivan, "An evaluation of uplift mapping languages," *International Journal of Web Information Systems*, vol. 13, pp. 00–00, Oct. 2017, doi: 10.1108/IJWIS-04-2017-0036.

[3] A. Ferrara and A. Nikolov, "Data Linking for the Semantic Web," *Int. J. Semantic Web Inf. Syst.*, vol. 7, pp. 46–76, Jan. 2011, doi: 10.4018/jswis.2011070103.

[4] C. Debruyne, B. Walshe, and D. O'Sullivan, "Towards a Project Centric Metadata Model and Lifecycle for Ontology Mapping Governance," in *Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services*, New York, NY, USA, 2015, vol. 1–10, p. Article 50. doi: 10.1145/2837185.2837201.

[5] P. Shvaiko and J. Euzenat, "Ontology Matching: State of the Art and Future Challenges," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 1, pp. 158–176, 2013, doi: 10.1109/TKDE.2011.253.

[6] N. Matentzoglu *et al.*, "A Simple Standard for Sharing Ontological Mappings (SSSOM)," *Database*, vol. 2022, p. baac035, Jan. 2022, doi: 10.1093/database/baac035.

[7] A. Dimou, T. De Nies, R. Verborgh, E. Mannens, P. Mechant, and R. Van de Walle, "Automated metadata generation for Linked Data generation and publishing workflows," in *LDOW2016*, 2016, pp. 1–10.

[8] H. Thomas, R. Brennan, and D. O'Sullivan, "Using the OM2R Meta-Data Model for Ontology Mapping Reuse for the Ontology Alignment Challenge: A Case Study," in *Proceedings of the 7th International Conference on Ontology Matching - Volume 946*, Aachen, DEU, 2012, pp. 61–72.

[9] S. Alzahrani, D O'Sullivan, "Towards a Unified Metadata Model for Semantic and Data Mappings", Ontology Matching workshop at ISWC 2022, October 2022, In Press.

Authorized licensed use limited to: TRINITY COLLEGE LIBRARY DUBLIN. Downloaded on April 21,2026 at 04:16:20 UTC from IEEE Xplore. Restrictions apply.

2024 IEEE 18th International Conference on Semantic Computing (ICSC) | 979-8-3503-8535-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICSC59802.2024.00058

MetaMap: Facilitating Mapping Reuse and Lifecycle Documentation

Sarah Alzahrani
Computer Science and Statistics
Trinity College Dublin
Dublin, Ireland
salzahra@tcd.ie

Declan O'Sullivan
ADAPT Centre, Computer Science and Statistics
Trinity College Dublin
Dublin, Ireland
declan.osullivan@tcd.ie

***Abstract*— In this research paper, we present 'MetaMap', a mapping annotation tool designed to serve an essential role in data and semantic mapping. MetaMap integrates key features for mapping annotation, incorporating metadata that meticulously captures the mapping lifecycle, with the hope of facilitating mapping reuse and mapping maintenance in subsequent mapping development phases. The tool's functionality unfolds in three core phases. Initially, users upload mappings and extract metadata. Then, users have the option to enrich their mappings with additional metadata fields, ensuring a comprehensive documentation from the lifecycle process of the associated mapping. Lastly, MetaMap empowers users with the flexibility to represent this metadata in various formats, such as named graphs or RDF-star.**

Keywords*— *Mapping, Alignment, Metadata Model, Knowledge Representation, Annotation.

I. Introduction

Declarative mappings are known as a valuable solution to enhance interoperability on the internet, where challenges stem from semantic heterogeneity and diverse data [1]. Heterogeneity and diversity is immediately evident when handling various datasets, ontologies, and publishing data online [2]. Declarative mappings offer a useful method for specifying the semantic relationships between different data models, addressing the intricacies of information exchange [1]. In the context of linked data and the semantic web, declarative mapping processes can be usefully categorized into three types: ontology mapping, interlinking, and uplift/downlift. Figure 1 provides an illustration of these mapping processes.

Each of the three mapping processes generates declarative mappings, albeit with distinct representations. Processed by engines, these mappings help with tasks like data transformation, alignment, and merging. Developing these mappings is complicated and requires expertise and time [3].

To save time and costs, it is important to reuse/repurpose previously made mappings and in general to manage and maintain the mappings given the evolution of datasets on either side of the mapping. Appropriate and detailed metadata associated with the mappings is clearly necessary to enable such development and management. While existing metadata associated with mappings is often limited and not easily queryable, prioritizing the mapping statement over the entire mapping development process, we address this common challenge in mapping metadata, as discussed in the related work section (Section II). This underscores the importance of a consistent documentation approach for mapping projects.

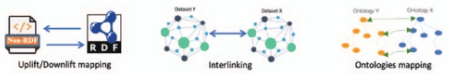


Figure 1 The Three categories of mappings

Our research focuses on a unified documentation model with detailed and accessible information with the aim of greatly benefiting knowledge engineers in their development, reuse and maintenance of mappings. The ultimate aim is to assist knowledge engineers in tasks such as finding, reusing, and evaluating the quality of mappings. As the foundation for the MetaMap tool, a mapping lifecycle has been formulated to accommodate all three types of mappings (presented earlier), along with a unified metadata model that comprehensively captures every aspect of mapping development—from its purpose and stakeholders to post-development, long-term management.

The proposed mapping lifecycle and metadata model was introduced in [4]. A survey experiment was conducted to collect feedback on the initial mapping metadata model from the semantic web and linked data community. The aim was to assess the relevance of the proposed metadata fields for two mapping-related tasks. Preliminary results, available in [5], indicate a generally positive response. In the experiment, most

979-8-3503-8535-9/24/$31.00 ©2024 IEEE
DOI 10.1109/ICSC59802.2024.00058

Authorized licensed use limited to: TRINITY COLLEGE LIBRARY DUBLIN. Downloaded on April 21,2026 at 04:18:47 UTC from IEEE Xplore. Restrictions apply.

of the proposed metadata fields were rated as relevant, and participants suggested additional useful metadata fields.

In the next section, we provide an overview of related work to position MetaMap within the broader context of existing research and tools in the field. Section III outlines the design and workflow of MetaMap, detailing its functions and user interfaces. The subsequent Section IV provides exploration of MetaMap's implementation, offering insights into technical considerations. In addition, Section V enriches the discussion with a practical case study. Finally, Section VI wraps up the paper by summarizing key points and discussing future research directions.

II. Related work

Several studies in the literature have introduced diverse metadata models and methods for enriching mappings with metadata [6][3][7]. However, many of these studies predominantly focus on metadata associated solely with the final mapping artefact. In contrast, a few notable ones, including [6] and [3], take a more holistic approach by considering the entire lifecycle of mapping development, extending beyond the final artefact. The OM2R project is particularly notable for being the first to comprehensively capture different stages of the ontology mapping process through its metadata model [6]. OM2R provides a specialized vocabulary for describing mappings while maintaining a clear distinction from the mappings themselves. Alternatively, the proposal in [3] introduces a novel ontology mapping lifecycle by adopting a project-centric approach. The corresponding metadata model, built as an ontology in OWL 2, extends the provenance ontology PROV-O[8] to comprehensively cover all mapping process steps.

In the realm of metadata generation, [9] presents an approach to automatically generate provenance and metadata information for linked data, implemented in the RML toolchain. Dimou et al. have pointed out that the type of metadata supported by their approach can be extended to accommodate other vocabularies. In the meantime, the RMLEditor [10] records the activities of document creation and editing, as well as when and by whom they were committed. Throughout their work, they have focused primarily on how to generate and represent metadata fields rather than what to capture, and as a consequence have captured limited metadata about mappings. In [7], a community standard called Simple Standard for Ontological Mappings (SSSOM) has been proposed as a means of sharing semantic mappings between information entities. In SSSOM, mapping sets and mappings are specified with schemas, data models, and metadata elements [11]. SSSOM, however, is limited in that it focuses on ontology mappings and the capture of metadata that describes the statements of the mappings and the provenance of the mappings without considering the entire lifecycle of the mapping.

As a final observation, it should be noted that previous studies often design metadata models focused on particular mapping types, as seen in [3] and [7], where the proposed models are specifically crafted for ontology mappings. Overall, there is a need for a unified metadata model applicable to any declarative mapping type, with extensibility options as needed. Furthermore, to enhance mapping discovery, reuse, and quality assessment, the unified metadata model should comprehensively capture all aspects of mapping development, ranging from the mapping purpose, and involved stakeholders to the long-term management post-development.

In comparison to the tools and studies in the state of the art, MetaMap is specifically designed for metadata enrichment for all three mapping types, considering the entire lifecycle of the mapping. Furthermore, in designing MetaMap the research explores different representations of metadata and how these metadata should be attached/associated with a mappingwhile maintaining the mapping structure.

III. Design And use

MetaMap is proposed as a web-based tool designed to equip knowledge engineers with an extensive and intricate metadata model, offering guidance throughout the process of annotation of declarative mappings. The tool's functionality extends the process to the reuse of mapping by empowering users to make informed decisions about the viability and appropriateness of reusing specific mappings.

The tool was developed in two distinct phases, each dedicated to creating, testing, and evaluating a specific component, with one component developed after the other. The initial components developed, namely the annotation and reuse components, were designed based on our research philosophy that emphasizes the role of metadata annotation in facilitating the reuse of mappings.

As discussed in Section I, the tool design began with the development of a mapping lifecycle capable of handling three different mapping types, accompanied by a comprehensive metadata model. Figure 2 illustrates the MetaMap solution design, highlighting four crucial aspects. Each of these aspects plays a vital role in guiding the tool's development. This section delves into a detailed discussion of each aspect, covering associated design decisions.

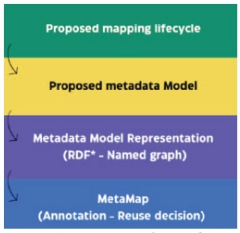


Figure 2 MetaMap solution design

A. Mapping lifecycle

The proposed lifecycle is crafted for ease of use by engineers with a general software engineering background, following industry practices to facilitate the learning process. Comprising

Authorized licensed use limited to: TRINITY COLLEGE LIBRARY DUBLIN. Downloaded on April 21,2026 at 04:18:47 UTC from IEEE Xplore. Restrictions apply.

six phases—Analysis, Mapping Design, Mapping Development, Mapping Testing, and Mapping Maintenance—the lifecycle is designed to be adaptable across the three declarative mapping types (ontology alignment, uplift/downlift, interlinking), with varying emphasis and application depending on the specific mapping type [4].

B. Metadata model

Metadata fields were introduced for each stage of the lifecycle to capture and document the activities and decisions made during the corresponding phase of mapping development (46 fields in total) [4][5]. Additionally, a set of competency questions was formulated for each lifecycle stage to facilitate the validation of the proposed metadata model. Subsequently, the model underwent formalization into a specification. Within this specification, each field in the metadata model is distinctly defined by its associated vocabulary term, expected value type, and a comprehensive description of its purpose or significance. Table 1 illustrates how the specification describes the fields by presenting one example[1].

Table 1 the metadata model specification

Phase	Field	Vocabulary term	Value	purpose
Analysis	Stakeholder's name	foaf:FirstName foaf:surname	Literal	To provide a human readable name for the stakeholders involved in the mapping project

Various RDF namespaces from well-known ontologies were employed to describe the metadata model. For instance, the FOAF vocabulary was utilized for capturing stakeholder information, encompassing their names and organizations. Additionally, the Dublin Core Metadata Initiative (DCMI) vocabulary was employed to detail aspects like the format, source, and creator of the input file. For metadata pertaining to the mapping itself, such as mapping requirements, domain, tools used, and mapping methods, custom namespace was created to encompass information about the mapping project This approach ensures that metadata fields align with existing RDF vocabularies, enhancing interoperability and widespread comprehension.

C. Metadata model representation

The representation and attachment of metadata to RDF-expressed mappings play a crucial role in storage efficiency, query performance, and usability [12]. Various RDF metadata approaches are found in the literature, such as standard reification, named graphs, singleton properties, and RDF-star.
Standard RDF reification involves describing a triple using four statements, known as reification, with vocabulary terms like rdf:Statement, rdf:predicate, and rdf:object. This method has drawbacks, including the lack of formal semantics and a significant increase in dataset size due to the need for four additional triples to represent one statement [12].

In an effort to overcome these drawbacks, [13] proposed the Singleton Property, offering formal semantics for statements about statements. This approach relies on predicates rather than subjects, amending the original predicate with a unique one to carry future metadata. Although it reduces the number of triples needed, it produces a large number of unique predicates [12].

Hartig & Thomson introduced RDFstar [14], extending the standard's syntax to allow nesting within the RDF graph. Using << and >> to enclose referenced triples, this approach results in smaller documents, facilitating faster data exchange and shorter SPARQL queries [15].

Named graphs [16] represent another annotation approach, where an RDF graph is named with a URI. The identifier of the named graph in the RDF graph can be used to make simple statements about the entire named graph.

[12] provides a comparison of RDF reification approaches—standard reification, singleton property, and RDFstar—based on triples, storage size, and query execution time. Results show differences, with Singleton Property reducing triples by about 40%, RDF-star by almost 70%, and variations in storage size and query execution time.

Additional benchmarking reported in [15] tested approaches on a Wikidata subset. Although Singleton Properties were excluded due to the production of numerous unique predicates, RDF-star demonstrated favorable results. A recent study in the area [17] explores the suitability of various knowledge graph representations for different consumer scenarios: knowledge exploration, systematic querying, and graph completion. The research compares representations—Standard Reification, N-Ary Relationships, Wikidata qualifiers, Named Graphs, and RDF-star—through user studies, SPARQL querying analysis, and graph embedding model experiments. The study offers valuable insights into the importance of representation choices and their impact on knowledge graph consumption scenarios, contributing to the ongoing discussion on optimal knowledge graph deployment.

In consideration of benchmarking contributions and recent studies, RDF-star and named graphs were chosen as representations for the proposed metadata model.

D. MetaMap

The initial phase of tool development outlined in this section primarily concentrated on crafting the annotation component. This component empowers users to upload mapping files and establish a metadata model associated with the mappings. Initially tailored to uplift mapping, the annotation component's functionality is expected to broaden progressively in future iterations, encompassing various mapping types and additional functionalities.
This section outlines the workflow, design, and output of the annotation component, specifically tailored for handling uplift

[1] http://surl.li/oknaa

Authorized licensed use limited to: TRINITY COLLEGE LIBRARY DUBLIN. Downloaded on April 21,2026 at 04:18:47 UTC from IEEE Xplore. Restrictions apply.

mappings. As illustrated in Figure 3, the annotation process within MetaMap typically comprises four key steps:

- Upload Mapping File: Initiate the process by uploading the mapping file into the system. Initially, the system considers only uplift mapping, with plans to incorporate all other mapping types in the future.
- Fill in Metadata Fields: This step involves two subfunctions. The first subfunction detects metadata from the mapping file, while the second requires users to provide additional metadata details.
- Annotate Mapping: This step involves selecting a mapping representation, choosing between RDF-star and named graphs, and annotating the mapping accordingly.
- Export result: After a review process, the annotated content can be viewed or downloaded to a result file.

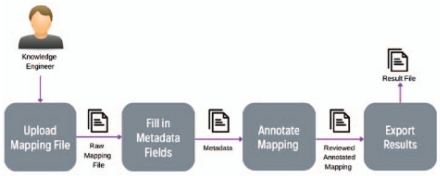


Figure 3 Typical workflow using MetaMap

The tool is publicly accessible [18], eliminating the requirement for user accounts. Users can effortlessly engage in mapping annotation and reuse directly through the web interface. Refer to Figure 4 for an illustration of the main website interface.

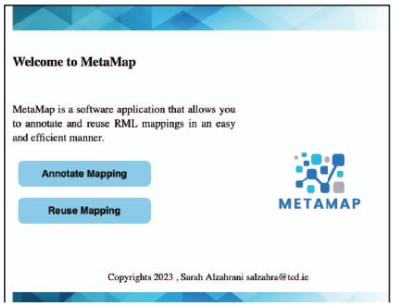


Figure 4 Home page of the MetaMap website

Following the previously outlined workflow, mapping annotation can be accomplished in four steps. The subsequent part of the section will demonstrate these steps with an example for annotating an RML mapping file.

In the example, the Attractions dataset, sourced from Data.gov.ie [19]—an Irish governmental open data portal—underwent uplift from CSV to RDF using the RMLEditor [10]. This process converts the data into RDF and generates mapping statements for subsequent annotation with MetaMap.

In the initial step, users are prompted to upload their mapping files. In this development phase, RML mapping files are accepted in Turtle format. With RDF Mapping Language (RML), heterogeneous data structures and serializations are mapped to RDF with customized rules. This mapping language is a superset of the W3C-standardized mapping language [R2RML], expanding its application and scope [20].

Upon uploading the mapping file, the tool automatically extracts recognized data, including the uploaded file name, mapped file names, and their formats—serving as the primary input for metadata, users proceed to another interface to add additional metadata fields. In this example a segment of the metadata model has been filled. As shown in Figure 5 These fields capture details related to the mapping development stage in the mapping lifecycle. Metadata includes the starting and ending dates of development, the tool utilized, as well as the mapping type and domain. After submitting metadata, users are prompted to review their entries and select their preferred presentation format. They have the option to choose between Named Graph, RDF-star, or both. Figure 5 provides a snapshot of the interface illustrating this step.

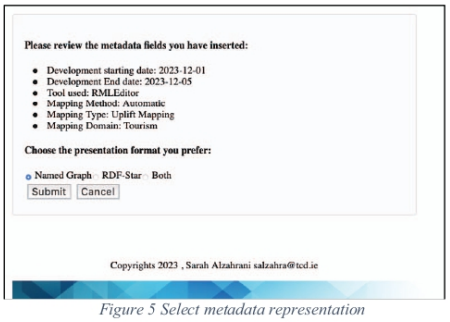


Figure 5 Select metadata representation

Upon a successful annotation process, users can either view or export the resulting file. Figure 6 displays the resulting file in the Named Graph format, while Figure 7 and 8 illustrates the result in RDF-Star.The RDF data model represents information in triples, comprising subject, predicate, and object. RDF triple stores accommodate multiple RDF graphs, allowing queries across different graphs concurrently. RDF datasets typically include a default graph (unnamed and potentially empty) and zero or more named graphs identified by IRIs, facilitating SPARQL queries for data retrieval [21]. To annotate the mapping using named graphs, the RML mapping statement file was treated as the default graph, followed by the creation of a new named graph (metadata graph) using SPARQL.

Authorized licensed use limited to: TRINITY COLLEGE LIBRARY DUBLIN. Downloaded on April 21,2026 at 04:18:47 UTC from IEEE Xplore. Restrictions apply.

```
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix ns1: <http://example.com/mappings/> .

ns1:subject ns1:endDate "2023-12-05" ;
    ns1:format "Named Graph" ;
    ns1:mappingDomain "Tourism" ;
    ns1:mappingMethod "Automatic" ;
    ns1:mappingType "Uplift Mapping" ;
    ns1:startDate "2023-12-01" ;
    ns1:toolUsed "RMLEditor" ;
    dcterms:creator "Data.gov.ie " ;
    dcterms:source "Data.gov.ie " ;
    foaf:familyName "Alzahrani" ;
    foaf:givenName "Sarah" ;
    foaf:organization "TCD" .
```

Figure 6 Metadata Graph

For enriching the mapping with metadata using RDF-star, a detailed examination of the RML statement structure was necessary. In Figure 7, lines 2 to 3, metadata describing the input file was seamlessly integrated with the logical source triple, adding dcterms:creator and dcterms:source to the entire logical source triple. Subsequently, in Figure 8, lines 2 to 11, the remaining metadata fields were incorporated with the complete TripleMap triple.

```
<<<http://example.com#LogicalSource2>
    dcterms:creator "Data.gov.ie";
    dcterms:source "Data.gov.ie">>
    rml:referenceFormulation ql:CSV;
    rml:source "Attraction.csv";
    a rml:LogicalSource.
```

Figure 7 logical source triple

```
<<<http://example.com#TriplesMap0>
    ns1:endDate "2023-12-05";
    ns1:format "Named Graph";
    ns1:mappingDomain "Tourism";
    ns1:mappingMethod "Automatic";
    ns1:mappingType "Uplift Mapping";
    ns1:startDate "2023-12-01";
    ns1:toolUsed "RMLEditor";
    foaf:familyName "Alzahrani";
    foaf:givenName "Sarah";
    foaf:organization "TCD">>
    rml:logicalSource <<http://example.com#LogicalSource2>>;
    a rr:TriplesMap;
    rr:predicateObjectMap ?pom;
    rr:subjectMap ?sm.
```

Figure 8 TriplesMap triple

IV. Implementation

MetaMap,was developed using Python[22], SPARQL[23], SPARQL-Star [14], and RML[20]. Python, powered by the Flask library [24], played a central role in creating the web application and executing SPARQL queries through RDF libraries [25] for mapping annotation and graph creation. The user-friendly graphical interface was crafted using HyperText Markup Language (HTML) [26] and Cascading Style Sheets (CSS) [27].In the initial stage of development, MetaMap supports declarative mappings of the uplift type (specifically RML initially) . The initial choice focus on using RML for transforming data into RDF was influenced by the necessity to accommodate various dataset formats. The code is available at the GitHub repository [2]which also contains a live demo website. In the second phase of development, the tool will expand its support to include the other types of mappings of interest in this research, that is interlinking and ontologies alignment.

V. Future work and Conclusion

Our initial research focused on the development of a mapping lifecycle and a foundational unified metadata model that will allow for the annotation and reuse of three major types of declarative mappings The initial research laid the groundwork for the development of MetaMap—an innovative software application designed for annotating and reusing mappings. As detailed in Section III, MetaMap comprises two fundamental components that will be systematically developed in two distinct phases. The initial phase focuses on the creation of the annotation component, which is expounded upon in this paper. Subsequently, in a subsequent phase as part of future work, the development will progress to include a reuse decision support component, as well as support for ontology and interlinking types of mappings. In the near term, the usability and functionality of the current tool will undergo comprehensive evaluation by users, ensuring a thorough assessment of its practical application and effectiveness. Taking a user-centric evaluation will provide valuable insights into the tool's performance, informing potential refinements and optimizations in subsequent iterations.

VI. Acknowledgment

The first author acknowledges financial support from Al Imam Mohammad Ibn Saud Islamic University and the Saudi government - represented by the Royal Embassy of Saudi Arabia - Cultural Bureau in Dublin. The second author is partially supported by the SFI ADAPT Research Centre (grant number 13/RC/2106_P2).

References

[1] M. Mao, “Ontology Mapping: An Information Retrieval and Interactive Activation Network Based Approach,” in *The Semantic Web*, K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, and P. Cudré-Mauroux, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 931–935.

[2] A. Ferrara and A. Nikolov, “Data Linking for the Semantic Web,” *Int. J. Semantic Web Inf. Syst.*, vol. 7, pp. 46–76, Jan. 2011, doi: 10.4018/jswis.2011070103.

[3] C. Debruyne, B. Walshe, and D. O’Sullivan, “Towards a Project Centric Metadata Model and Lifecycle for Ontology Mapping Governance,” in *Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services*, in iiWAS ’15, vol. 1–10. New York,

[2] https://github.com/sarah-alzahrani/MetaMap

Authorized licensed use limited to: TRINITY COLLEGE LIBRARY DUBLIN. Downloaded on April 21,2026 at 04:18:47 UTC from IEEE Xplore. Restrictions apply.

NY, USA: Association for Computing Machinery, 2015, p. Article 50. doi: 10.1145/2837185.2837201.
[4] S. Alzahrani and D. Sullivan, "Towards a Unified Metadata Model for Semantic and Data Mappings," in *Ontology Matching workshop at ISWC 2022*, 2022.
[5] S. Alzahrani and D. O'Sullivan, "What metadata is needed for semantic and data mappings?," in *2023 IEEE 17th International Conference on Semantic Computing (ICSC)*, 2023, pp. 312–316. doi: 10.1109/ICSC56153.2023.00064.
[6] H. Thomas, R. Brennan, and D. O'Sullivan, "Using the OM2R Meta-Data Model for Ontology Mapping Reuse for the Ontology Alignment Challenge: A Case Study," in *Proceedings of the 7th International Conference on Ontology Matching - Volume 946*, in OM'12. Aachen, DEU: CEUR-WS.org, 2012, pp. 61–72.
[7] N. Matentzoglu *et al.*, "A Simple Standard for Sharing Ontological Mappings (SSSOM)," *Database*, vol. 2022, p. baac035, Jan. 2022, doi: 10.1093/database/baac035.
[8] "PROV-O: The PROV Ontology." https://www.w3.org/TR/prov-o/ (accessed Dec. 14, 2023).
[9] A. Dimou, T. De Nies, R. Verborgh, E. Mannens, P. Mechant, and R. Van de Walle, "Automated metadata generation for Linked Data generation and publishing workflows," in *LDOW2016*, CEUR-WS. org, 2016, pp. 1–10.
[10] "RMLEditor | RML.io." https://rml.io/tools/rmleditor/ (accessed Mar. 29, 2023).
[11] N. Matentzoglu *et al.*, "A Simple Standard for Ontological Mappings 2022: Updates of data model and outlook".
[12] F. Orlandi, D. Graux, and D. O'Sullivan, "Benchmarking RDF Metadata Representations: Reification, Singleton Property and RDF," in *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, Laguna Hills, CA, USA: IEEE, Jan. 2021, pp. 233–240. doi: 10.1109/ICSC50631.2021.00049.
[13] V. Nguyen, O. Bodenreider, and A. Sheth, "Don't like RDF Reification? Making Statements about Statements Using Singleton Property," in *Proceedings of the 23rd International Conference on World Wide Web*, in WWW '14. New York, NY, USA: Association for Computing Machinery, 2014, pp. 759–770. doi: 10.1145/2566486.2567973.
[14] O. Hartig, "Foundations of RDF* and SPARQL*:(An alternative approach to statement-level metadata in RDF)," in *AMW 2017 11th Alberto Mendelzon International Workshop on Foundations of Data Management and the Web, Montevideo, Uruguay, June 7-9, 2017.*, Juan Reutter, Divesh Srivastava, 2017.
[15] "What is RDF-Star? - Ontotext." https://www.ontotext.com/knowledgehub/fundamentals/what-is-rdf-star/ (accessed Apr. 26, 2023).
[16] J. J. Carroll, C. Bizer, P. Hayes, and P. Stickler, "Named Graphs, Provenance and Trust," in *Proceedings of the 14th International Conference on World Wide Web*, in WWW '05. New York, NY, USA: Association for Computing Machinery, 2005, pp. 613–622. doi: 10.1145/1060745.1060835.
[17] A. Iglesias-Molina, K. Ahrabian, F. Ilievski, J. Pujara, and O. Corcho, "Comparison of Knowledge Graph Representations for Consumer Scenarios," in *International Semantic Web Conference*, Springer, 2023, pp. 271–289.
[18] "MetaMap." https://metamap--sarah-alzahrani.repl.co/ (accessed Dec. 19, 2023).
[19] "Data.gov.ie." https://data.gov.ie/ (accessed Mar. 29, 2023).
[20] "RDF Mapping Language (RML)." https://rml.io/specs/rml/ (accessed Mar. 29, 2023).
[21] M. Barbieri, "Working with RDF Database Named Graphs," *agnos.ai*, Aug. 19, 2022. https://medium.com/agnos-ai/working-with-rdf-database-named-graphs-a5ddab447e91 (accessed Apr. 05, 2023).
[22] "Welcome to Python.org," *Python.org*, Dec. 10, 2023. https://www.python.org/ (accessed Dec. 11, 2023).
[23] "SPARQL 1.1 Query Language." https://www.w3.org/TR/sparql11-query/ (accessed Dec. 11, 2023).
[24] "Welcome to Flask — Flask Documentation (3.0.x)." https://flask.palletsprojects.com/en/3.0.x/ (accessed Dec. 11, 2023).
[25] "rdflib 7.0.0 — rdflib 7.0.0 documentation." https://rdflib.readthedocs.io/en/stable/ (accessed Dec. 11, 2023).
[26] "W3C HTML." https://www.w3.org/html/ (accessed Dec. 11, 2023).
[27] "Cascading Style Sheets." https://www.w3.org/Style/CSS/ (accessed Dec. 11, 2023).

Authorized licensed use limited to: TRINITY COLLEGE LIBRARY DUBLIN. Downloaded on April 21,2026 at 04:18:47 UTC from IEEE Xplore. Restrictions apply.

MetaSEMAP: Usability Evaluation for Metadata Annotation Across the Mapping Lifecycle*

Sarah Alzahrani[1,2,*,†], Declan O'Sullivan[3,†]

[1]School of Computer Science and Statistics, Trinity College Dublin, Ireland
[2] Imam Mohammad Ibn Saud Islamic University (IMSIU), Saudi Arabia
[3]ADAPT Center, Trinity College Dublin, Ireland

Abstract

Comprehensive and well-structured metadata annotation and documentation are essential for supporting the reuse and maintenance of declarative mappings throughout their lifecycle. This paper presents a usability evaluation of MetaSEMAP, a tool designed to facilitate the annotation of declarative mappings, including uplift mappings, ontology alignment, and interlinking. While MetaSEMAP and its underlying metadata model are still under active development, the focus of this work is on evaluating how users interact with the tool and interpret the metadata concepts it introduces. The evaluation investigates MetaSEMAP's ability to support metadata annotation using real-world scenarios such as the Virtual Record Treasury of Ireland. Participants provided feedback on the tool's usability and their preferences for metadata representation, including RDF-star and Named Graph. The results reveal both strengths and areas for improvement, offering valuable insights for the development of more effective mapping annotation tools. This work supports efforts to improve interoperability and sustainability in mapping practices, with alignment to FAIR principles as a longer-term goal.

Keywords

Metadata Annotation, Mapping Annotation, Declarative Mappings, Mapping Lifecycle, RDF-star, Named Graph, FAIR Principles, Usability

1. Introduction

Declarative mappings are essential for interoperability across diverse data ecosystems, helping to resolve issues related to semantic heterogeneity and varying data structures [1]. Such diverse data ecosystems range from the deployment of semantic web technologies into infrastructure-type approaches, e.g., event-based networking [2], right through to complex multi-domain applications, e.g., building information management [3]. These mappings typically fall into three categories: ontology alignment, uplift mapping, and interlinking. However, while declarative mappings provide a useful means of linking and transforming data, challenges arise in managing, understanding, and reusing these mappings. Their lifecycles are complex, involving multiple stakeholders, evolving formats, and shifting requirements, all of which complicate their reuse and long-term sustainability [1].

The importance of metadata in addressing these challenges cannot be overstated. Metadata provides the critical context needed to understand the purpose, domain, contributors, and technical characteristics of declarative mappings. However, existing approaches often lack standardized metadata, offer limited queryability, and provide only partial coverage of the mapping lifecycle. This makes essential tasks such as reuse, maintenance, versioning, and governance difficult to carry out effectively.

To address these limitations, we introduce MetaSEMAP, a metadata-driven tool designed to help users annotate declarative mappings in a structured and consistent way. MetaSEMAP is built upon a metadata model that formalizes the lifecycle of mapping development, comprising five distinct phases: Analysis,

First International Workshop on Users and Knowledge Graphs (UKG), co-located with SEMANTiCS'25: International Conference on Semantic Systems, September 3–5, 2025, Vienna, Austria

*Corresponding author.
†These authors contributed equally.
salzahra@tcd.ie (S. Alzahrani); declan.osullivan@tcd.ie (D. O'Sullivan)
https://www.scss.tcd.ie/Declan.OSullivan/ (D. O'Sullivan)
0000-0003-2169-7006 (S. Alzahrani); 0000-0001-7770-2117 (D. O'Sullivan)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

CEUR-WS.org/Vol-4064/UKG-paper2.pdf

Design, Development, Testing, and Maintenance. Each phase is associated with specific metadata fields relevant to its role in the lifecycle. The proposed metadata model not only improves consistency and completeness but also enables automated validation and machine-readable documentation, supporting the FAIR principles, especially in terms of interoperability and reusability.

MetaSEMAP provides a simple, web-based interface through which users can annotate uplift mapping, ontology alignment, or interlinking mappings using a guided form based on the proposed metadata model. The tool supports both RDF-star and Named Graph representations to offer flexibility in metadata expression. To evaluate the effectiveness of MetaSEMAP, we conducted a usability study with 50 participants, assessing perceived usability, task completion time, and collecting qualitative feedback on the user experience.

While MetaSEMAP and the underlying metadata model are still under development, this paper does not aim to evaluate their completeness or technical implementation. Instead, the primary contribution is an exploratory usability evaluation, focusing on how users engage with the tool and interpret the metadata concepts it introduces. The results offer insights into user needs and guide future improvements to both the tool and the model.

The remainder of the paper is structured as follows. Section 2 reviews related work on metadata for declarative mappings and annotation tools. Section 3 introduces the different types of declarative mappings and outlines the proposed mapping lifecycle. Section 4 presents the metadata model that supports annotation across lifecycle phases. Section 5 describes the MetaSEMAP tool and its implementation. Section 6 outlines the evaluation design and the metrics used to assess usability and performance. Section 7 presents the results and insights from the study, including quantitative findings, qualitative feedback, and representation preferences. Section 8 discusses the broader implications for metadata design and usability. Finally, Section 9 concludes the paper with a summary of key findings and future directions.

2. Related Work

Efforts to manage and share metadata for declarative mappings often focus on specific mapping types or particular stages of the mapping lifecycle, rather than offering comprehensive solutions that support documentation, traceability, and reusability across all phases.

For ontology alignment, Thomas et al. proposed OM2R, a metadata model capturing the full lifecycle of ontology mappings to support their reuse and management [4]. While effective in the alignment context, it was not designed to generalize across other mapping types such as interlinking or uplift. The Simple Standard for Sharing Ontological Mappings (SSSOM) is a community-developed specification for representing ontology and terminology alignments [5]. It defines a set of rich metadata fields such as predicate, confidence score, and justification to describe individual mapping assertions. While SSSOM effectively supports documenting the results of the alignment process, it offers limited coverage of metadata related to the broader mapping lifecycle, such as stakeholder roles, tool usage workflows, or quality assurance procedures. Although some of these aspects may be accommodated through community-driven extensions or through integration with complementary vocabularies such as PROV-O, they are not explicitly addressed in the core SSSOM specification.

In the interlinking domain, tools like Silk [6] and LIMES [7] support link discovery across datasets. Although they are effective for generating RDF links, these tools do not incorporate features for documenting metadata about the linking methodology, stakeholder decisions, or validation strategies.

A recent initiative, FAIR-IMPACT, has advocated for improving the FAIRness of mapping documentation. It promotes the use of structured metadata, persistent identifiers, and semantic enrichment to enhance the findability and reusability of mapping artefacts across their lifecycle [8, 9]. As part of this effort, FAIR-IMPACT recommends the use of the Simple Standard for Sharing Ontological Mappings (SSSOM) and has proposed extensions to support additional metadata fields [8]. While these extensions improve the ability to describe mappings with provenance and justifications, the focus remains largely on output-level assertions. The proposed updates still do not fully capture the broader mapping lifecycle,

such as stakeholder involvement, design rationale, tool usage, or validation steps. As a result, manual annotation support and integration into end-to-end mapping workflows remain limited.

Complementing these initiatives, Toledo et al. proposed RMLdoc, a tool that generates human-readable documentation for RML mapping files [10]. While effective in increasing the transparency of uplift mappings, RMLdoc does not extend to other mapping types such as ontology alignment or interlinking, nor does it support full lifecycle documentation.

MetaSEMAP addresses these limitations by providing a unified, implementation-level metadata model and annotation tool that supports ontology alignment, interlinking, and uplift mappings. It enables users to document activities across all phases of the mapping lifecycle, such as stakeholder involvement, design decisions, testing processes, and maintenance practices.

3. Declarative Mappings: Types and Lifecycle

Declarative mappings are central to achieving interoperability on the Semantic Web and, as previously noted, can be grouped into three categories: ontology alignment, interlinking, and uplift mapping. These processes support the transformation, linking, and semantic integration of heterogeneous data. Figure 1 provides an illustrative representation of these mapping types.

This study addresses all three. Uplift mappings used to convert data into RDF, e.g. [11]. Interlinking identifies relationships between entities across datasets [12], while ontology alignment establishes correspondences between concepts in different ontologies to enable semantic interoperability [13].

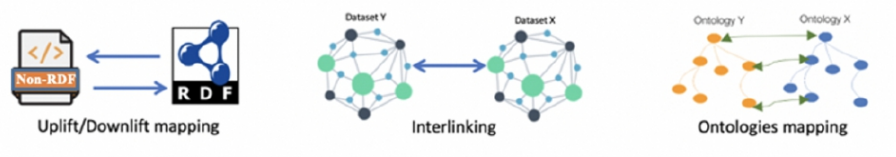


Figure 1: The three categories of mappings.

To structure these activities, we propose a mapping lifecycle inspired by prior models [14, 4, 1], with the addition of a dedicated testing phase. The resulting lifecycle includes five stages: Analysis, Design, Development, Testing, and Maintenance. This framework is designed to be applicable across all mapping types, with phase relevance varying by context.

4. Metadata Model

Building on the proposed mapping lifecycle, we developed a metadata model to document key activities across all phases. Metadata fields were designed to capture decisions and actions specific to each phase of the lifecycle [15]. Table 1 summarizes the structure of the model, aligning metadata fields with corresponding lifecycle stages.

To validate the model's structure and the relevance of its fields, a survey was conducted with participants from the Semantic Web and Linked Data community [16]. The goal was to assess the applicability of the proposed metadata for two mapping-related tasks. Results showed strong agreement with the selected fields, and participants also suggested additional useful metadata elements.

In the initial tool implementation (see next section), the model reused widely adopted vocabularies such as FOAF (for stakeholder roles), and DCMI (for metadata about inputs like source, format, and creator), along with a custom namespace for domain-specific fields (e.g., requirements, tools).

A more complete ontology is under development[1], incorporating standard vocabularies including

[1]Project repositry: https://github.com/SarahAlzahranitcd/MetaSEMAP-Metadata

Table 1
Proposed Metadata Fields Aligned with Mapping Lifecycle Phases

Lifecycle Phase	Key Metadata Fields
Analysis	Stakeholder details (URI, name, background, role, organization); mapping purpose (requirements, type, domain, assumptions, technical needs, risks); input descriptions (URI, name, source, type, creator, format)
Design	Final design decisions, justifications, and anticipated quality metrics
Development	Mapping details (URI, name, start/end date, tools used, mapping method, algorithm, format)
Testing	Testing metadata (type, timestamp, results)
Maintenance	Versioning information (publisher name, source, version number, date)

PROV-O [17], DQV [18], and more recent vocabularies such as MQV [19] to formally describe provenance, quality assessment, and validation metadata. A detailed specification describing each metadata field and its usage is also provided in the same repository. This ontology will be integrated into the next version of the tool and evaluated with knowledge engineers during the second development phase of this project.

5. The Initial MetaSEMAP Tool

MetaSEMAP evolved from MetaMap [20], initially focused on uplift mappings, and was renamed to reflect broader support and avoid naming conflicts. MetaSEMAP is a web tool that supports structured annotation of mappings. The homepage (Figure 2) offers two workflows: annotate new or reuse existing mappings.

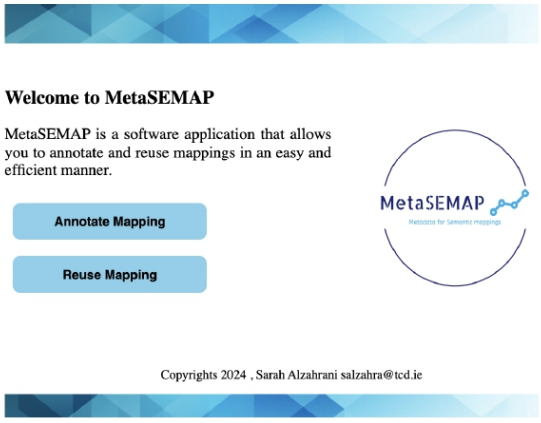


Figure 2: MetaSEMAP main interface

The annotation process in MetaSEMAP unfolds through four key steps. First,Users begin by uploading an RDF mapping file (Figure 3).

Next, users complete structured metadata fields aligned with the mapping lifecycle. These fields prompt users to document key aspects of the mapping, such as its purpose, input file, development method, and associated stakeholders. As shown in Figure 4, the form is organized into sections that correspond to each lifecycle phase. Some fields are scenario-specific and optional, while others are required to ensure metadata completeness and consistency.



Figure 3: MetaSEMAP Step 1: Upload and preview of mapping file before annotation.

To assist users in understanding each metadata category, contextual help was provided via tooltips. As illustrated in Figure 5, hovering over a question mark icon revealed a short explanation about the purpose and expected input for the associated metadata section. This feature was especially useful for users and helped ensure more consistent and accurate annotations.

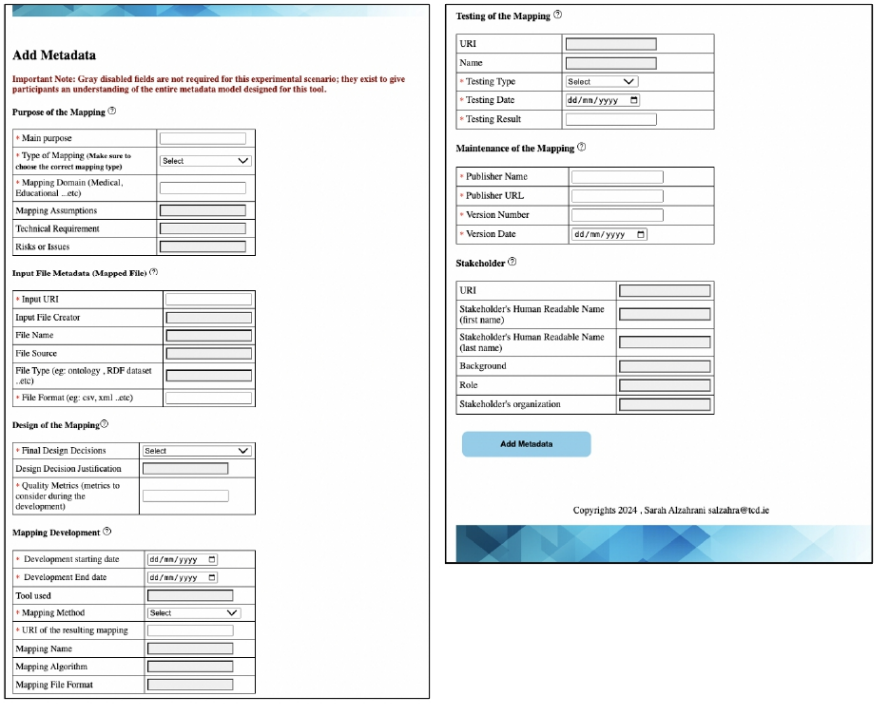


Figure 4: MetaSEMAP metadata annotation interface.

After completing the metadata form, users are presented with a review screen (Figure 6) summarizing all the fields they have entered. This confirmation step was designed to promote careful review and reduce input errors before submission. Once the confirmation is done, the tool generates machine-

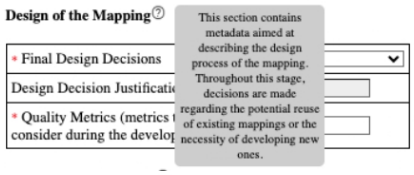


Figure 5: Tooltip example: Metadata field explanations appeared when hovering over category labels.

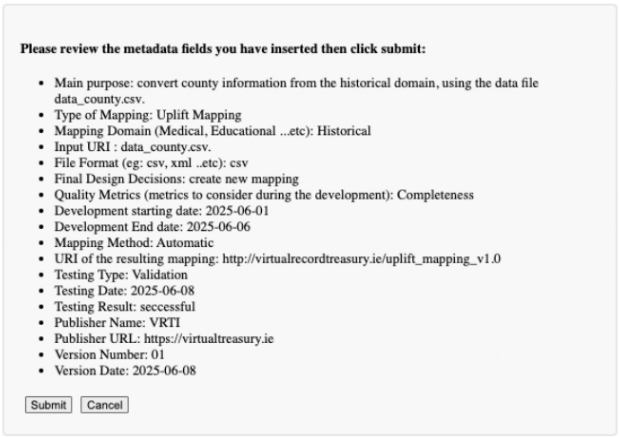


Figure 6: Review screen.

readable metadata in either RDF-star or Named Graph. Users can preview and download the output for further use or publication. Figure 7 illustrates the final interface that users see after annotation, showing the confirmation message and access to RDF-star and Named Graph outputs.

6. Evaluation Design

To evaluate usability and the effectiveness of the metadata model for declarative mappings, we designed a controlled user evaluation of the initial MetaSEMAP implementation, guided by the following research questions:

- **RQ1:** How usable is the MetaSEMAP tool for annotating different types of declarative mappings?
- **RQ2:** How effectively can users complete metadata annotation tasks using the proposed metadata model?
- **RQ3:** Which metadata representation (Named Graph or RDF-star) do users find easier to read and interpret?

6.1. Participants

Fifty MSc students enrolled in a knowledge and data engineering course participated in the user evaluation study. They had introductory knowledge of mapping processes but varied in technical expertise. The study was approved by the institutional ethics committee, and participation was voluntary, anonymous, and online. Participants used the public MetaSEMAP website (MetaSEMAP Experiment) asynchronously.

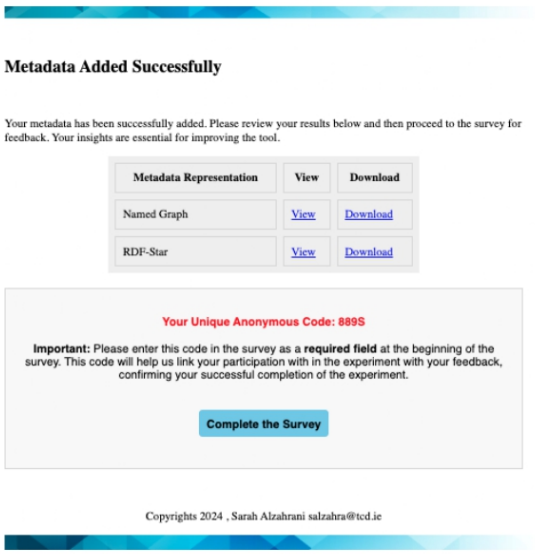


Figure 7: MetaSEMAP interface showing post-annotation confirmation.

6.2. Tasks and Scenarios

Each participant was randomly assigned one of three scenarios representing a different declarative mapping type: uplift mapping, ontology alignment, or interlinking. These scenarios were grounded in the Virtual Record Treasury of Ireland (VRTI) knowledge graph [21]. VRTI is a historical knowledge graph designed to digitally reconstruct Ireland's archival heritage, offering structured, domain-rich data suitable for metadata annotation experiments. Each scenario included a mapping file and contextual metadata. Participants were instructed to:

1. Read the scenario and download the provided mapping file.
2. Upload the file to MetaSEMAP.
3. Annotate the mapping using the tool's metadata fields, referencing the scenario.
4. Review the generated metadata in both RDF-star and Named Graph representations.
5. Complete a survey reflecting on usability and representation preference.

The scenario summaries are listed in Table 2.

6.3. Evaluation Metrics

To assess the study goals, we employed a mix of quantitative and qualitative metrics. These included:

- *System usability and satisfaction* measured via the Post-Study System Usability Questionnaire (PSSUQ), based on a 1–7 Likert scale survey.[2]
- *Task completion time* tracked from start to metadata submission.
- *Representation preference* captured through survey feedback.
- *Thematic feedback* from open-ended responses.

The anonymized survey responses and the mapping files annotated by participants during the study are available in the MetaSEMAP project repository 1^0.

[2]PSSUQ Survey: https://docs.google.com/forms/d/e/1FAIpQLSdRhBPxzCQwrQyi3zmMPdP-EeON7zqFAG5tR6aoSd0UnXuMXQ/viewform

Table 2
Summary of Scenarios in the Usability Experiment

Mapping Type	Summary of Scenario
Uplift Mapping	Convert county information from historical data into RDF triples. The data file `data_county.csv` was used to map county IDs, names, and geographic details to enrich the VRTI Knowledge Graph.
Ontology Alignment	Align person entities (e.g., historical figures) between the VRTI Knowledge Graph and external datasets. Features such as relationships, affiliations, and roles were matched using the file `person_alignment.rdf`.
Interlinking	Connect Irish historical figures to their corresponding Wikidata entries using `owl:sameAs` relationships. The project enriched VRTI records with biographical and contextual data using the file `interlinking.rdf`.

7. Results and Discussion

7.1. System Usability and User Satisfaction

System usability was assessed using a modified version of the Post-Study System Usability Questionnaire (PSSUQ), adapted to include a subset of questions inspired by the System Usability Scale (SUS). The final instrument consisted of 14 items covering system usefulness, information quality, and interface quality, rated on a 1–7 Likert scale (lower is better). Negatively phrased questions (e.g., Q8, Q9, Q11, Q13) were reverse-coded prior to analysis, in line with standard usability survey practices. Out of 50 participants, 46 completed the questionnaire. The overall average score was 2.49, indicating good usability. Around 70% rated the tool positively, citing ease of use and effective support for metadata annotation.

System Usefulness scored 2.4, reflecting user confidence and successful error recovery. Information Quality received a score of 3.4, suggesting moderate satisfaction, particularly with error message clarity. Interface Quality was rated at 2.3, pointing to a generally intuitive and consistent layout. Table 3 summarizes these scores and their associated questions.

Table 3
Average Scores by PSSUQ Category and Associated Questions

Category	Avg. Score	Questions Used
System Usefulness	2.4	Q4, Q5, Q6, Q7
Information Quality	3.4	Q1, Q2, Q8, Q9
Interface Quality	2.3	Q3, Q10, Q11, Q13
Overall Usability	2.49	All of the above + Q12, Q14

7.2. Task Completion Time

Participants completed annotation tasks based on one of three mapping types. The initial average completion time was 44 minutes. To ensure the results reflected realistic task durations, outliers were excluded using empirical thresholds: submissions under 5 minutes (likely rushed or incomplete) and those exceeding 200 minutes (suggesting prolonged interruptions or technical issues). After removing these outliers, the adjusted averages were: 32 minutes for Uplift Mapping, 39 minutes for Ontology Alignment, and 33 minutes for Interlinking.

7.3. Preferred Metadata Representation

Participants reviewed both RDF-star and Named Graph outputs generated from their annotations. A majority (74.4%) expressed a preference for Named Graphs. In contrast, 25.6% preferred RDF-star.

One potential influencing factor was the representation setup: RDF-star annotations were included in the same file as the mapping (see Listing 1), and participants may have found them harder to interpret without prior exposure to the syntax. Named Graph metadata, on the other hand, was presented in a separate file (see Listing 2). This physical separation may have enhanced the perceived readability and organization of Named Graphs. Additionally, inconsistencies in prefix usage between the two representations may have further affected participant preferences. These aspects will be addressed in future iterations to ensure more balanced comparison conditions.

Listings 1 and 2 show an example of RDF-star and Named Graph annotations for an uplift mapping, where the `rr:TriplesMap` triple was the subject of annotation. For ontology alignment, RDF-star annotations targeted the `align:Alignment` instance URI, while for interlinking, the RDF-star subject was the SPARQL query resource used to perform linking between VRTI and Wikidata. In all cases, the Named Graph representation followed a consistent pattern, where metadata was written in a separate file using a `metag:subject` placeholder and reused the same structured fields as RDF-star.

```
@prefix b2022: <https://ont.virtualtreasury.ie/ontology#> .
@prefix cidoc: <http://erlangen-crm.org/current/> .
@prefix geo: <http://www.opengis.net/ont/geosparql#> .
@prefix ql: <http://semweb.mmlab.be/ns/ql#> .
@prefix rml: <http://semweb.mmlab.be/ns/rml#> .
@prefix rr: <http://www.w3.org/ns/r2rml#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://example.com/ns##COUNTY> a rr:TriplesMap ;
   rml:logicalSource [
      rml:query "SELECT␣VRTI_ID,␣NAME_TAG␣FROM␣'data_county.csv'" ;
      rml:referenceFormulation ql:CSV ;
      rml:source "data_county.csv"] ;
   rr:subjectMap [
      rr:template "https://kg.virtualtreasury.ie/place/county/{NAME_TAG}/{VRTI_ID}" ;
      rr:class cidoc:E53_Place] ;
   rr:predicateObjectMap [
      rr:predicate geo:hasCentroid ;
      rr:objectMap [ rr:template "https://kg.virtualtreasury.ie/place/county/centroid/{NAME_TAG}/{
          VRTI_ID}" ]] .

<< <http://example.com/ns##COUNTY> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> rr:TriplesMap >>
   <http://example.com/metag/purpose> "convert␣county␣information␣from␣the␣historical␣domain,␣using␣
       the␣data␣file␣data_county.csv." ;
   <http://example.com/metag/mappingType> "Uplift␣Mapping" ;
   <http://example.com/metag/mappingDomain> "Historical" ;
   <http://example.com/metag/startDate> "2025-06-01" ;
   <http://example.com/metag/endDate> "2025-06-06" ;
   <http://example.com/metag/mappingMethod> "Automatic" ;
   <http://example.com/metag/testingType> "Validation" ;
   <http://example.com/metag/testingResult> "seccessful" ;
   <http://example.com/metag/versionNumber> "01" ;
   <http://example.com/metag/versionDateTime> "2025-06-08" ;
   <http://xmlns.com/foaf/0.1/organization> "VRTI" .
```

Listing 1: RDF-star Annotation with Mapping Statement (shortened)

```
@prefix dcmi: <http://purl.org/dc/terms/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix metag: <http://example.com/metag/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

metag:subject
   metag:purpose "convert␣county␣information␣from␣the␣historical␣domain,␣using␣the␣data␣file␣
       data_county.csv." ;
```

```
    metag:mappingType "Uplift␣Mapping" ;
    metag:mappingDomain "Historical" ;
    metag:startDate "2025-06-01"^^xsd:date ;
    metag:endDate "2025-06-06"^^xsd:date ;
    metag:mappingMethod "Automatic" ;
    metag:testingType "Validation" ;
    metag:testingResult "seccessful" ;
    metag:versionNumber "01" ;
    metag:versionDateTime "2025-06-08"^^xsd:date ;
    foaf:organization "VRTI" .
```

Listing 2: Named Graph Annotation (shortened)

Note: These examples are shortened to highlight the structure and differences between RDF-star and Named Graph metadata representations. Complete examples are available in the project repository 1[0].

7.4. Metadata Annotation Observations

In addition to usability metrics, we conducted a light-touch review of the metadata submissions to observe common patterns and challenges in annotation. Participants generally completed structured fields (e.g., start dates, URIs, publisher names) with high accuracy, suggesting that these fields were intuitive and well-supported by the interface. However, free-text fields such as *purpose* and *qualityMetrics* exhibited greater variability in both content and specificity.

Some responses conveyed the intended meaning, but lacked precision or used overly general terms. For example, several participants filled in the *mapping purpose* field with generic phrases like 'to describe the data' or 'link information', rather than specifying the actual goal of the project or the user need that motivated the mapping. More conceptual fields, such as distinguishing between manual and automatic mapping methods, also introduced occasional ambiguity.

These patterns indicate that while the tool successfully supports basic annotation tasks, certain metadata elements may benefit from additional in-tool guidance. Features such as example values, tooltips, or predefined templates could help users provide more precise and consistent entries, particularly in scenarios involving complex or abstract metadata concepts.

7.5. Qualitative Feedback

Open-ended responses were analyzed to extract key themes:

- **Ease of Use and Simplicity:** MetaSEMAP was generally described as intuitive and non-technical.
- **Error Handling:** Users requested clearer error messages and better field validation, especially for URIs.
- **Field Guidance:** Participants asked for more in-tool help (e.g., tooltips or explanations) to understand metadata fields.
- **Interface Improvements:** Suggestions included larger input fields, better text wrapping, and persistent scenario visibility.
- **Additional Features:** Users proposed templates, collaboration support, and enhanced consistency checks.

User feedback emphasized priorities for improvement, such as enhanced error messages, better field-level guidance, and improved layout design. These findings will inform future enhancements of MetaSEMAP to support both usability and clarity in metadata annotation workflows.

8. Discussion

The evaluation results show that MetaSEMAP is generally usable across different mapping types, with high scores for system usefulness and interface quality. However, areas such as error message clarity

and guidance for free-text fields remain improvement priorities. Participants strongly preferred Named Graphs over RDF-star, which appears to be influenced by presentation structure and familiarity. Named Graphs were delivered in a separate file and followed consistent prefixing, which likely enhanced readability and perceived structure. The current annotation strategy focused on a project-level scope, (e.g., alignment file, SPARQL query, or RML mapping). Although this proved sufficient in our controlled scenarios, it may be less reliable or too coarse-grained in larger use cases, particularly when mappings involve many modular components. Future work will explore more fine-grained annotation. In addition to usability, the collected metadata can support mapping reuse, quality assurance, and contextual understanding by other users. Fields such as purpose, assumptions, and stakeholder roles help assess whether a mapping suits a new use case, while testing metadata and versioning support maintenance and trust. Although some metadata elements require effort to complete, they provide long-term value in ensuring mappings are interpretable, reusable, and aligned with FAIR principles. Overall, the evaluation confirms the value of lifecycle-aware metadata annotation and highlights practical lessons for designing usable tooling in this space.

9. Conclusion and Next Steps

This paper's primary contribution is the usability evaluation of MetaSEMAP, rather than a full validation of the metadata model or complete tool functionality, both of which are still under development. Nevertheless, the feedback offers concrete guidance for improving metadata design, user support, and metadata representation in future iterations of the tool.

An open question emerging from this work is whether structured metadata remains necessary when large language models are able to generate or infer declarative mappings independently. On the other hand, incorporating contextual metadata about the mapping project such as domain, purpose and assumptions into LLM prompts may improve the quality, relevance, and explainability of the generated mappings.

Future work will investigate how structured metadata, when used as contextual input in prompt-based mapping generation, affects the quality and reliability of LLM outputs. This direction aims to assess how human-in-the-loop metadata practices and AI-assisted approaches can work together to support sustainable knowledge graph construction.

Acknowledgments

The first author acknowledges financial support from Al Imam Mohammad Ibn Saud Islamic University and the Saudi government — represented by the Royal Embassy of Saudi Arabia — Cultural Bureau in Dublin. The second author is partially supported by the SFI ADAPT Research Centre (grant number 13/RC/2106_P2).

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT (GPT-4) and Grammarly for the purposes of grammar improvements. All AI-generated content was thoroughly reviewed, edited, and validated by the author, who takes full responsibility for the final manuscript and all its content.

References

[1] C. Debruyne, B. Walshe, D. O'Sullivan, Towards a project centric metadata model and lifecycle for ontology mapping governance, in: Proc. 17th Int. Conf. on Info. Integration and Web-based Applications & Services (iiWAS '15), ACM, 2015, p. Article 50. doi:10.1145/2837185.2837201.

[2] J. Keeney, D. Roblek, D. Jones, D. Lewis, D. O'Sullivan, Extending siena to support more expressive and flexible subscriptions, in: Proceedings of the 2nd International Conference on Distributed Event-Based Systems (DEBS), 2008.
[3] K. McGlinn, R. Brennan, C. Debruyne, A. Meehan, L. McNerney, E. Clinton, P. Kelly, D. O'Sullivan, Publishing authoritative geospatial data to support interlinking of building information models, Automation in Construction 124 (2021) 1–15.
[4] H. Thomas, R. Brennan, D. O'Sullivan, Moom – a prototype framework for management of ontology mappings, in: Proc. Int. Conf. on Adv. Information Networking and Applications (AINA), 2011, pp. 548–555. doi:10.1109/AINA.2011.55.
[5] N. Matentzoglu, J. P. Balhoff, S. M. Bello, M. H. Brush, S. Carbon, M. Courtot, A. D. Diehl, D. Dooley, H. J. Drabkin, M. A. Harris, M. Haendel, C. J. Mungall, D. A. Natale, D. Osumi-Sutherland, N. Perry, K. Shefchek, B. Smith, N. Vasilevsky, N. Vasilevsky, J. A. McMurry, A simple standard for sharing ontological mappings (sssom), Database 2022 (2022) baac035. URL: https://doi.org/10.1093/database/baac035. doi:10.1093/database/baac035.
[6] R. Volz, et al., Silk: A link discovery framework for the web of data, in: Proceedings of the International Semantic Web Conference, 2009.
[7] A.-C. N. Ngomo, S. Auer, Limes - a time-efficient approach for large-scale link discovery on the web of data, in: Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI), 2011, pp. 2312–2317.
[8] Z. Martinková, P. Escobar, M. Galkin, V. Rodríguez-Doncel, et al., D4.5: Guidelines for enriching mappings and crosswalks with FAIR-enabling metadata, Technical Report, FAIR-IMPACT Project, 2024. URL: https://zenodo.org/records/15111167, zenodo, DOI:10.5281/zenodo.15111167.
[9] FAIR-IMPACT Consortium, Semantic artefact mappings, 2024. URL: https://fair-impact.eu/semantic-artefact-mappings, accessed June 2025.
[10] J. Toledo, A. Iglesias-Molina, D. Chaves-Fraga, D. Garijo, Rmldoc: Documenting mapping rules for knowledge graph construction, in: ESWC24 Posters and Demos Track, 2024.
[11] Rdf mapping language (rml), https://rml.io/specs/rml/, 2023. Accessed: Mar. 29, 2023.
[12] A. Ferrara, A. Nikolov, F. Scharffe, Data linking for the semantic web, Int. J. Semantic Web Inf. Syst. 7 (2011) 46–76.
[13] J. Euzenat, An api for ontology alignment, in: S. A. McIlraith, D. Plexousakis, F. van Harmelen (Eds.), The Semantic Web – ISWC 2004, Springer Berlin Heidelberg, 2004, pp. 698–712.
[14] D. O'Sullivan, V. Wade, D. Lewis, Understanding as we roam, IEEE Internet Comput. 11 (2007) 26–33. doi:10.1109/MIC.2007.50.
[15] S. Alzahrani, D. O'Sullivan, Towards a unified metadata model for semantic and data mappings, in: Ontology Matching Workshop at ISWC, 2022.
[16] S. Alzahrani, D. O'Sullivan, What metadata is needed for semantic and data mappings?, in: 2023 IEEE 17th International Conference on Semantic Computing (ICSC), 2023, pp. 312–316. doi:10.1109/ICSC56153.2023.00064.
[17] L. Moreau, et al., Prov-o: The provenance ontology, 2013. URL: https://www.w3.org/TR/prov-o/.
[18] C. L. F. Rocha, et al., Data quality vocabulary (dqv), 2015. URL: https://www.w3.org/TR/vocab-dqv/.
[19] M. Randles, et al., Mapping quality vocabulary (mqv): A domain-specific model for capturing metadata and provenance relating to mapping quality assessment, refinement, and validation, Journal of Semantic Web 15 (2021) 305–323. URL: https://doi.org/10.1007/s11220-021-00325-4. doi:10.1007/s11220-021-00325-4.
[20] S. Alzahrani, D. O'Sullivan, Metamap: Facilitating mapping reuse and lifecycle documentation, in: 2024 IEEE 18th International Conference on Semantic Computing (ICSC), 2024, pp. 323–328. doi:10.1109/ICSC59802.2024.00058.
[21] A. Randles, L. McKenna, L. Kilgallon, B. Yaman, P. Crooks, D. O'Sullivan, The knowledge graph explorer for the virtual record treasury of ireland, in: VOILA Workshop - International Semantic Web Conference (ISWC), CEUR, 2024, pp. 1–15.

Guiding LLM Generated Mappings with Lifecycle-Based Metadata: An Early Evaluation

Sarah Alzahrani[1,2,*], Declan O'Sullivan[3]

[1]School of Computer Science and Statistics, Trinity College Dublin, Ireland
[2] Imam Mohammad Ibn Saud Islamic University (IMSIU), Saudi Arabia
[3]ADAPT Center, Trinity College Dublin, Ireland

Abstract

Large Language Models (LLMs) are increasingly used to automate knowledge engineering tasks such as generating RDF mappings. While promising, their outputs often lack semantic precision, syntactic correctness, and contextual metadata. This paper investigates whether structured metadata aligned with the mapping lifecycle can improve the quality and reusability of LLM-generated mappings. We present a metadata model that covers key phases of the mapping process and integrate it into the MetaSEMAP tool to support context-aware prompting. Using real-world uplift scenarios, we compare RML outputs generated from unguided prompts with those informed by lifecycle metadata. Our initial findings show that guided prompts consistently produce syntactically valid, semantically rich, and FAIR-aligned mappings. These results highlight the potential of structured metadata to guide LLMs toward generating higher-quality and reusable semantic artifacts in knowledge graph construction.

Keywords

Metadata, Declarative mappings, LLMs, Mapping lifecycle, Knowledge graphs, Context engineering

1. Introduction

Declarative mappings like RML[1] are essential for transforming structured data into RDF for use in knowledge graphs. However, creating such mappings is a technically demanding task, often requiring expertise in both syntax and domain-specific ontologies [1]. Recently, large language models (LLMs) have shown promise in assisting with this task by generating knowledge graphs and mappings from natural language descriptions [2, 3, 4]. Despite their capabilities, LLMs often struggle with producing outputs that are semantically accurate, syntactically valid, and compliant with established standards [5], and human-in-the-loop validation is often recommended to ensure quality and reliability.

This paper explores whether incorporating structured metadata into LLM prompts can improve the quality of generated mappings. This approach aligns with recent work in context engineering, which focuses on enriching prompts with structured and task-specific context to enhance the reliability and relevance of LLM outputs [6]. We focus on a real-world scenario involving the creation of mappings to uplift data from Ireland's open data portal[2].and demonstrate how metadata guidance, based on a mapping lifecycle model, influences the outputs generated by gpt-3.5-turbo (via OpenAI API).

The remainder of this paper is structured as follows: we review related work on declarative mappings, metadata standards, and the role of LLMs in semantic data generation. We then introduce our lifecycle-based metadata model and its integration into the MetaSEMAP prompting interface. Next, we present the experiment design and results comparing guided and unguided prompt outputs. Finally, we conclude the paper and outline future research directions.

The Second Workshop on Knowledge Graphs and Neurosymbolic AI (KG-NeSy), co-located with SEMANTiCS'25: International Conference on Semantic Systems, September 3–5, 2025, Vienna, Austria
salzahra@tcd.ie (S. Alzahrani); declan.osullivan@tcd.ie (D. O'Sullivan)
https://www.scss.tcd.ie/Declan.OSullivan/ (D. O'Sullivan)
0000-0003-2169-7006 (S. Alzahrani); 0000-0001-7770-2117 (D. O'Sullivan)
© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[1]RML (RDF Mapping Language) is a declarative language for mapping heterogeneous data to RDF. See: http://rml.io/spec.html
[2]Ireland's open data portal: https://data.gov.ie

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

CEUR-WS.org/Vol-4064/kgnesy-paper4.pdf

2. Background and Related Work

2.1. Declarative Mappings and Metadata Need

Declarative mapping languages such as R2RML and its extension RML enable structured transformation of heterogeneous data sources into RDF for use in knowledge graphs [1]. These languages define how data is semantically uplifted without requiring custom code. RML, in particular, supports various input formats like CSV, JSON, and XML using the RML vocabulary in combination with query languages such as XPath and JSONPath. Despite its expressiveness, authoring RML mappings remains a technically demanding task. Tools like YARRRML [7] and documentation frameworks such as RMLDoc [8] aim to simplify this process by offering user-friendly syntaxes or structured documentation. However, the lack of embedded metadata to describe the purpose, provenance, and context of mappings limits their interpretability and reuse.

2.2. Metadata Standards for Mappings

To enhance transparency and reusability, several metadata standards have been proposed. General-purpose vocabularies like Dublin Core and DCAT provide terms for describing datasets and publishing context. More specialized efforts, such as SSSOM [9], focus on metadata for ontology alignments, while MQV [10] supports the assessment of mapping quality. Recent initiatives like FAIR-IMPACT [11] proposed metadata and best practices for describing mappings as FAIR Digital Objects, aiming to improve their findability, accessibility, and interoperability. These efforts offer valuable guidance on core metadata elements such as provenance, mapping purpose, and versioning, but remain largely conceptual and do not yet support a comprehensive, lifecycle-based framework. As a result, a unified metadata model that cover the full mapping lifecycle including design, generation, documentation, and reuse is still lacking. To address this, we propose a structured lifecycle-based metadata model that captures key elements across five phases of mapping activities. This model, introduced and evaluated in our earlier work [12, 13], is further detailed later in the paper.

2.3. Large Language Models in Semantic Mapping

Large Language Models (LLMs), such as GPT-3 and GPT-4, have demonstrated the ability to generate RDF triples, SPARQL queries, and declarative mappings such as RML from natural language inputs [3]. This capability presents new opportunities for simplifying the knowledge graph construction process, particularly for non-experts. However, while promising, the outputs of LLMs often lack semantic precision and completeness. Studies have shown that LLM-generated mappings frequently omit critical components such as input declarations, namespace prefixes, and join conditions, and they rarely include metadata supporting provenance, reuse, or quality assessment [14, 3, 2]. Moreover, these mappings may include hallucinated classes or properties that do not exist in the target ontology, making them unreliable for use.

To address these challenges, we propose a structured metadata-driven approach that guides LLMs using context aligned with the mapping lifecycle. By injecting metadata such as the mapping's purpose, data source characteristics, design decisions, and provenance, we provide the model with domain-specific constraints and intentions that help shape the generated outputs. This approach improves the accuracy, interpretability, and reusability of the resulting mappings by grounding generation in explicit and verifiable context.

This aligns with emerging practices in context engineering [6], where structured metadata and formal vocabularies help constrain and contextualize LLM behavior to produce more accurate and consistent outputs. In our case, the metadata model acts as a source of prompt structure, supporting reproducibility and enhancing trust in LLM-assisted mapping generation.

3. Lifecycle Metadata Model

To support high-quality and reusable semantic mappings, we proposed a lifecycle-based metadata model that organizes metadata across five key phases: analysis, design, development, testing, and maintenance. This model was first introduced in our earlier work [12] and later validated through a community study with Semantic Web practitioners [13]. It is implemented in the MetaMap tool [15], enabling structured metadata capture through a guided interface.

The model includes fields for describing the mapping purpose, input sources, stakeholders, design decisions, tooling, validation outcomes, and publishing context. Table 1 presents a summary of the key metadata fields grouped by lifecycle phase. The full specification is publicly available on GitHub [16].

In this paper, we extend the model's use beyond documentation by incorporating selected fields into LLM prompts. This approach evaluates whether structured metadata can guide LLMs to generate more accurate, standards-compliant, and reusable RML mappings.

Table 1
Proposed Metadata Fields Aligned with Mapping Lifecycle Phases

Lifecycle Phase	Key Metadata Fields
Analysis	Stakeholder details (URI, name, role, organization); mapping purpose (requirements, domain, assumptions, risks); input description (URI, name, source, type, creator, format)
Design	Final design decisions, justifications, expected quality metrics
Development	Mapping process metadata (mapping URI, tools used, method, algorithm, start/end date, format)
Testing	Testing metadata (type, timestamp, test results)
Maintenance	Versioning and publishing information (publisher name, source, version number, version date)

4. Methodology

4.1. Experimental Setup

To evaluate whether lifecycle-based metadata improves the quality of LLM-generated RML mappings, we designed an experiment comparing two prompting strategies: (1) unguided prompts with only a task description, and (2) guided prompts augmented with structured metadata. We used `gpt-3.5-turbo` via the OpenAI Python SDK v1.0 (`chat.completions.create`) to generate RML outputs. All prompts were submitted through a controlled interface (Figure 1) to ensure consistent model behavior and prompt formatting across both guided and unguided conditions.

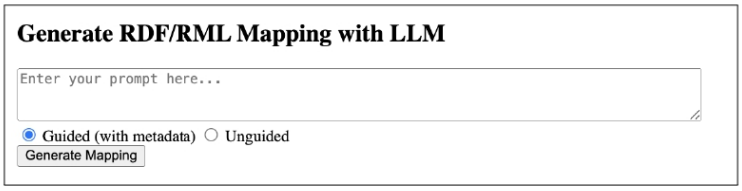


Figure 1: The interface to generate RML outputs.

4.2. Mapping Scenarios

We evaluate our approach using three real-world scenarios derived from publicly available datasets on Ireland's open data portal[2]. Full dataset links and mapping files are available in our GitHub repository[3]. In Scenario 1 (S1), we use the *Counties, National Statutory Boundaries, 2019* dataset to convert a CSV file containing administrative boundary data for Irish counties into RDF. Although the dataset includes multiple attributes such as object IDs, area size, Irish names, and shape geometry, only a relevant subset was selected for the purpose of the experiment.

In Scenario 2 (S2), we use the *G0421, Population per NUTS 3 Region* dataset to convert a JSON file containing population data for Irish regions into RDF. Although the dataset includes various statistical fields such as region codes, units of measurement, and percentage values, only a relevant subset was selected for the purpose of the experiment. The mapping focused on region identifiers, population counts, and temporal attributes to construct RDF triples that reflect regional population distributions in a given year.

In Scenario 3 (S3), we use the *CSO Electoral Divisions, National Statistical Boundaries, 2022* dataset to convert a CSV file containing administrative division data for electoral districts in Ireland into RDF. Although the dataset includes various attributes such as object IDs, Irish and English names, shape geometry, and codes from multiple authorities, only a relevant subset was selected for the purpose of the experiment. The mapping focused on modeling spatial hierarchy and containment relationships. Each electoral division was assigned a hierarchical URI based on its `ED_ID` and linked to its parent county and province using a custom property `ex:containedIn`.

4.3. Prompt Design

To compare the effects of metadata guidance, each mapping scenario was submitted to gpt-3.5-turbo using two types of prompts: *unguided* and *guided*. The unguided prompt included only a task description, while the guided prompt was enriched with structured metadata drawn from our lifecycle-based model, which defines 37 fields across five phases. We selected 17 fields most relevant to prompt-based mapping generation, focusing on descriptors from the *Analysis* and *Design* phases (e.g., *Purpose*, *Mapping Type*, *Mapping Domain*, *Input Description*, *Final Design Decisions*, *Justification*, and *Quality Metrics*). We also included contextual publishing and output fields from the *Maintenance* phase such as *Publisher Source*, *Version Number*, *Version Date*, *Output Format*, and *Output Syntax*, which clarify expected structure and encourage standards compliance. These fields were chosen based on their impact on design-time decisions and feedback from previous user studies [13] showing their relevance to mapping quality and reuse.

Fields from the *Development* and *Testing* phases (e.g., *Tools*, *Mapping Algorithm*, *Testing Result*) were excluded as they relate to post-generation steps. Similarly, fields with limited influence on generative quality (e.g., *Risks*, *Stakeholder Background*) were omitted. The selected metadata was embedded in natural language form to simulate structured context engineering and guide the LLM in producing syntactically valid, semantically rich, and reusable RML mappings. Table 2 shows an example of both prompt types used in Scenario 1 (S1). The full prompt sets for Scenarios 2 and 3 are available in our GitHub repository[3].

5. Evaluation

Each mapping was assessed based on three key dimensions. *Correctness* refers to whether the output is syntactically valid RML and can be parsed without errors. *Structure awareness* evaluates how well the mapping captures the structure of the input data, including the correct use of `rml:logicalSource`, `rml:referenceFormulation`, and iterators. *Semantic quality* considers whether the mapping uses meaningful classes, properties, and URIs aligned with domain semantics and whether it includes relevant metadata to support reuse. Across all three scenarios, guided prompts produced mappings that

[3]Project repository: https://github.com/sarah-alzahrani/LLM.

Table 2
Example of Guided and Unguided Prompt for Scenario 1

Prompt Type	Prompt Text
Unguided	Generate an RML mapping to convert a CSV file of Irish county boundaries into RDF. Each row contains a county ID, a county name, and geometry coordinates.
Guided	Generate an RML mapping to convert a CSV file of Irish county boundaries into RDF. Each row contains a county ID, a county name, and geometry coordinates. **Metadata:** **Stakeholder Name:** Ordnance Survey Ireland, **Role:** Data provider **Purpose:** Publish administrative boundaries as linked geodata **Mapping Type:** CSV to RDF (RML) **Mapping Domain:** Geospatial / Administrative **Input Description:** CSV file with county ID, names, and geometry **Input Name:** counties.csv , **Input Format:** CSV **Final Design Decisions:** Use `rdfs:label`, `schema:latitude`, `schema:longitude`, and `schema:addressRegion` to represent the attributes. **Justification:** Align with schema.org and LinkedGeoData best practices **Quality Metrics:** All 32 counties covered, valid spatial attributes **Publisher Source:** https://data.gov.ie/dataset/counties-national-statutory-boundaries-20191 **Version Number:** 1.0 , **Version Date/Time:** 2025-06-04 **Output Format:** Turtle (.ttl), **Output Syntax:** RML

were consistently more correct, structurally aware, and semantically richer than those from unguided prompts. In particular, all guided prompts correctly used `rml:logicalSource`, specifying input type and reference formulation (e.g., `ql:CSV` or `ql:JSONPath`), and included `iterator` when handling JSON inputs. In contrast, unguided prompts often defaulted to `rr:logicalTable`, which is valid R2RML syntax but insufficient for non-tabular or nested data. This distinction is important because `logicalSource` is the required mechanism in RML for handling diverse data sources, and its absence leads to incorrect or incomplete mappings, especially for formats like JSON. For example:

- **Scenario 1 (Counties – CSV)**: The unguided mapping used `rr:logicalTable`, ignoring CSV-specific reference formulation. The guided version used `rml:logicalSource` with `ql:CSV`, along with appropriate classes and coordinate properties.
- **Scenario 2 (Population – JSON)**: The unguided prompt omitted both the `iterator` and JSON path references, resulting in an unusable mapping. The guided prompt correctly used `ql:JSONPath`, specified the `iterator`, and aligned the schema with population data standards.
- **Scenario 3 (Electoral Divisions – CSV)**: The unguided output lacked hierarchical URI structure and containment logic. The guided version constructed meaningful URIs using region and division identifiers, and modeled geographic relationships with custom and standard vocabularies.

These results highlight how metadata-enriched prompting helps LLMs handle different mapping complexities by injecting contextual knowledge into generation. Guided prompts also produced reusable outputs that included metadata blocks aligned with FAIR and provenance principles, which were entirely missing from unguided versions. Overall, our initial evaluation shows that context-aware prompting grounded in lifecycle metadata improves both the technical correctness and semantic value of LLM-generated mappings, supporting better reuse, validation, and documentation. While this early evaluation offers qualitative insights, future work will expand the analysis with quantitative scoring, completeness metrics, and broader use case coverage to more systematically validate the approach.

6. Conclusion and Future Work

This study investigated the effect of lifecycle-based metadata guidance on the quality of RML mappings generated by LLM. By comparing guided and unguided prompts, we observed that metadata-enriched prompting significantly improves the syntactic accuracy, semantic richness, and standards compliance of LLM-generated mappings. Guided prompts led to the consistent use of appropriate vocabularies, correct handling of input formats, and the inclusion of metadata blocks that enhance provenance and reuse. These initial findings support a context-aware approach that combines the generative flexibility of LLMs with structured metadata to improve the reliability, interpretability, and reusability of semantic outputs. In future work, we plan to extend this evaluation framework to additional mapping types, compare performance across different LLMs, as well as investigate whether prompting LLMs to generate declarative mappings (like RML) results in more reusable outputs than directly generating RDF triples. Our initial findings highlight the value of structured metadata not just for documentation but as a guide in prompt-based knowledge graph construction. As LLMs become embedded in semantic workflows, the community must consider not only what outputs they generate but also how those outputs are guided, contextualized, and made reusable for others.

7. Acknowledgments

The first author acknowledges financial support from Al Imam Mohammad Ibn Saud Islamic University and the Saudi government — represented by the Royal Embassy of Saudi Arabia — Cultural Bureau in Dublin. The second author is partially supported by the SFI ADAPT Research Centre (grant number 13/RC/2106_P2).

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT (GPT-4) and Grammarly for the purposes of grammar improvements. All AI-generated content was thoroughly reviewed, edited, and validated by the author, who takes full responsibility for the final manuscript and all its content.

References

[1] A. Dimou, M. V. Sande, P. Colpaert, E. Mannens, R. Verborgh, R. V. de Walle, Rml: A generic language for integrated rdf mappings of heterogeneous data, in: Proceedings of the 7th Workshop on Linked Data on the Web (LDOW), 2014.

[2] M. Freund, R. Dorsch, S. Schmid, A. Harth, Mapping by example: Towards an rml mapping reverse engineering pipeline, in: ESWC 2025 Workshop on Knowledge Graph Construction (KGCW), 2025.

[3] M. Hofer, J. Frey, E. Rahm, Towards self-configuring knowledge graph construction pipelines using llms – a case study with rml, in: Proceedings of the 6th International Workshop on Knowledge Graph Construction (KGCW@ESWC), 2024.

[4] B. Zhang, H. Soh, Extract, define, canonicalize: An llm-based framework for knowledge graph construction, arXiv preprint arXiv:2404.03868 (2024). URL: https://arxiv.org/abs/2404.03868. doi:10.48550/arXiv.2404.03868, proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing.

[5] V. K. Kommineni, B. König-Ries, S. Samuel, From human experts to machines: An llm supported approach to ontology and knowledge graph construction, arXiv preprint arXiv:2402.11155 (2024). URL: https://arxiv.org/abs/2403.08345.

[6] A. Volkov, What is context engineering?, 2024. URL: https://blog.getzep.com/what-is-context-engineering/.

[7] P. Heyvaert, A. Dimou, R. Verborgh, Yarrrml: Yet another rml rendering language, in: Proceedings of the 1st International Workshop on Knowledge Graph Construction (KGC 2018), volume 2160 of *CEUR Workshop Proceedings*, 2018.

[8] P. Heyvaert, A. Dimou, R. Verborgh, Rmldoc: Generating documentation for declarative mapping rules, in: Proceedings of the 22nd International Semantic Web Conference (ISWC 2023), volume 14265 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 169–185. doi:10.1007/978-3-031-41393-4_11.

[9] N. Matentzoglu, et al., A simple standard for sharing ontological mappings (sssom), Database 2022 (2022) baac035. doi:10.1093/database/baac035.

[10] A. Randles, A. C. Junior, D. O'Sullivan, A vocabulary for describing mapping quality assessment, refinement and validation, in: 2021 IEEE 15th International Conference on Semantic Computing (ICSC), 2021, pp. 425–430. doi:10.1109/ICSC50631.2021.00076.

[11] Y. Le Franc, N. Grau, N. Juty, G. Mejias, P. Reed, P. Ramezani, M. Poveda-Villalón, D. Garijo, C. Goble, R. van Horik, D4.5 - Guidelines and methodology to create, document and share mappings and crosswalks (V1.1), Technical Report, FAIR-IMPACT, 2025. URL: https://doi.org/10.5281/zenodo.15111167. doi:10.5281/zenodo.15111167.

[12] S. Alzahrani, D. Sullivan, Towards a unified metadata model for semantic and data mappings, in: Ontology Matching workshop at ISWC 2022, 2022.

[13] S. Alzahrani, D. O'Sullivan, What metadata is needed for semantic and data mappings?, in: 2023 IEEE 17th International Conference on Semantic Computing (ICSC), 2023, pp. 312–316. doi:10.1109/ICSC56153.2023.00064.

[14] M. Parciak, B. Vandevoort, F. Neven, L. Peeters, S. Vansummeren, Schema matching with large language models: An experimental study, in: Proceedings of the VLDB 2024 Workshops, CEUR-WS, 2024.

[15] S. Alzahrani, D. O'Sullivan, Metamap: Facilitating mapping reuse and lifecycle documentation, in: 2024 IEEE 18th International Conference on Semantic Computing (ICSC), 2024, pp. 323–328. doi:10.1109/ICSC59802.2024.00058.

[16] S. Alzahrani, Metasemap metadata model, 2025. URL: https://github.com/SarahAlzahranitcd/MetaSEMAP-Metadata.

MMV: A Lifecycle-Based Ontology for Declarative Mapping projects

Sarah Alzahrani
Computer Science and Statistics
Trinity College Dublin
Dublin, Ireland.
salzahra@tcd.ie

Declan O'Sullivan
ADAPT Centre, Computer Science and Statistics
Trinity College Dublin)
Dublin, Ireland.
declan.osullivan@tcd.ie

***Abstract*— Declarative mappings support semantic interoperability across heterogeneous data sources, enabling tasks such as ontology alignment, data uplift, and interlinking. However, mapping projects are often documented only through their final artefacts, with limited structured metadata describing their context, development process, and evolution. This paper presents MMV (Mapping Metadata Vocabulary), a lifecycle-based ontology designed to represent metadata for declarative mapping projects across different mapping types. MMV captures contextual and procedural metadata spanning analysis, design, development, testing, and maintenance, while reusing established Semantic Web vocabularies. The paper describes the design of MMV, illustrates its instantiation, and outlines a planned expert evaluation.**

Keywords*— *declarative mappings, mapping metadata, ontology design, mapping lifecycle, semantic interoperability

I. Introduction

Declarative mappings play a central role in enabling semantic interoperability across heterogeneous data sources, supporting tasks such as data transformation, ontology alignment, and interlinking [1]. In this work, a *mapping* is understood as a declarative specification that defines correspondences or transformations between heterogeneous data models, schemas, or ontologies. Such mappings are typically expressed using formal rule-based languages and are executed or evaluated by mapping engines as part of knowledge graph construction, data integration, and semantic evaluation workflows.[1]

Despite their importance, mapping projects are commonly documented primarily through their final artefacts (e.g., mapping files or alignment outputs), with limited structured metadata describing how and why those artefacts were produced [2]. Contextual information such as mapping purpose, domain assumptions, tool configurations, validation procedures, quality measurements, and maintenance history is often scattered across informal documentation or omitted entirely. In addition, mappings increasingly are complex spanning the entire range of complex correspondences patterns [3] and incorporating functions for transformation logic [4]. This lack of systematic metadata hinders traceability, complicates comparison across projects, and limits the reliable reuse of mapping artefacts in new contexts.

These challenges are particularly evident in large-scale and collaborative initiatives, where mappings are developed, evaluated, and revised by multiple stakeholders over time. Without lifecycle-oriented metadata, it becomes difficult to analyze how mappings evolve across versions and to support transparency and governance.

This paper builds on a body of our prior work that has examined mapping metadata from complementary perspectives. Previous studies investigated (i) the metadata needs of mapping practitioners through a community survey, (ii) the design and implementation of MetaSEMAP, a web-based tool for capturing mapping metadata, (iii) the usability of structured metadata annotation through usability user studies, and (iv) the role of large language models (LLMs) in generating declarative mappings with and without structured metadata guidance [5],[6],[7],[8]. Collectively, these works indicates that lifecycle-oriented metadata is both needed and practically capturable, but also highlights the absence of a clearly articulated, reusable ontology to structure such metadata consistently across mapping types and tools.

In this paper, we address this gap by describing the Mapping Metadata Vocabulary (MMV) as a formal ontology. MMV is a lifecycle-based ontology designed to represent metadata about declarative mapping projects across analysis, design, development, testing, and maintenance. MMV is independent of specific mapping languages or execution frameworks and is intended to support a range of declarative mapping types, including ontology alignment, uplift mappings, and interlinking.MMV is instantiated through **MetaSEMAP** [6.7], which provides a structured interface for creating MMV-compliant metadata without requiring users to directly model OWL.

To assess the adequacy of MMV as an ontology, in this paper we outline a planned expert evaluation focusing on conceptual clarity, coverage, and alignment with established ontology engineering best practices. By formalising lifecycle-oriented metadata in a reusable ontology, this work aims to support more systematic documentation, analysis, and reuse of declarative mapping projects.

II. Motivating Mapping projects and metadata needs

A. Motivating Example

As an illustrative case, we use the OAEI 2025 Anatomy track to exemplify one instance of the broader challenges associated with documenting declarative mapping projects.

The Ontology Alignment Evaluation Initiative (OAEI) has organized annual evaluation campaigns since 2004 to assess ontology matching systems across multiple tracks, datasets, and evaluation settings [9]. For each track, participating systems generate alignment artefacts that are published as independent RDF alignment files, together with aggregated evaluation results based on standard metrics such as precision, recall, F-measure, recall+, and coherence.

In the OAEI 2025 campaign [10] Anatomy track for example, systems generate alignments between the Adult Mouse Anatomy ontology and the NCI Thesaurus. The resulting raw alignment files are made publicly available per system and per track, while details about the experimental setting such as the evaluation platform used (e.g., MELT, SEALS, or developer-provided execution), configuration assumptions, coherence checks, and known limitations are described separately in accompanying documentation [11].

Although this setup enables quantitative comparison of systems within the campaign, the alignment artefacts themselves are published without structured, machine-readable metadata describing how they were produced, under which assumptions, or how they relate to other artefacts across executions or versions. As a result, it is difficult to systematically trace the provenance of individual alignments, compare artefacts produced under different conditions, or assess their suitability for reuse beyond the original evaluation context.

This situation reflects a broader challenge in declarative mapping projects: mapping artefacts are commonly shared as standalone outputs, detached from lifecycle-oriented metadata describing their creation, testing, and evolution. Addressing this gap motivates the need for a structured metadata vocabulary for declarative mapping projects.

B. Metadata Requirements Across the Lifecycle

The motivating example illustrates that declarative mapping projects encompass more than the production of final mapping artefacts. Across their lifecycle, such projects involve a series of activities ranging from initial analysis to long-term maintenance each of which introduces contextual factors that influence how mappings are interpreted, developed and evaluated.

At early stages, metadata is required to describe the purpose and scope of a mapping project, including intended use, domain assumptions, and requirements that guide modelling decisions. During design and development, metadata is needed to document implementation choices, such as selected approaches, tools, configurations, and dependencies. Testing and evaluation phases introduce the need for metadata describing validation activities, including evaluation criteria, quality metrics, and assessment outcomes. Finally, maintenance phases require metadata that captures versioning, derivation, updates, and responsible stakeholders, enabling traceability across previous mapping versions.

In practice, this information is often fragmented across informal documentation or omitted entirely. The absence of a unified, lifecycle-oriented metadata model makes it difficult to systematically compare mapping projects, trace how mapping artefacts evolve, or assess their suitability for reuse across settings.

Based on these observations, we identify the following requirements for mapping metadata:

- R1 (Project-level context): Metadata must describe mapping projects as first-class entities, including purpose, scope, stakeholders, and temporal boundaries.
- R2 (Lifecycle activities): Metadata must capture distinct lifecycle activities (analysis, design, development, testing, maintenance) and their relationships to mapping artefacts.
- R3 (Artefact documentation): Metadata must describe mapping artefacts, including type, language, version, and derivation relationships.

These requirements motivate the design of a lifecycle-based ontology for mapping metadata, which we introduce in the following section.

III. MMV Ontology Design and structure

This section presents MMV (Mapping Metadata Vocabulary), a lifecycle-based ontology for representing metadata about declarative mapping projects. MMV is designed to capture not only mapping artefacts, but also the activities, assumptions, evaluation procedures, and maintenance actions that shape their creation. The ontology is intended to support documentation, traceability, and reuse across different declarative mapping types, including uplift mappings, ontology alignment, and interlinking.

A. Ontology Engineering Approach and Scope

The development of MMV followed established ontology engineering practices, drawing in particular on principles from lightweight ontology development methodologies such as the Linked Open Terms (LOT) approach [12] These methodologies emphasis requirement-driven design, reuse of existing vocabularies, and iterative refinement informed by evaluation. MMV adopts these principles to ensure conceptual clarity, modularity, and interoperability.

MMV is scoped to the project level, modelling mapping as evolving processes rather than isolated files. The ontology does not prescribe a specific mapping language or execution framework, allowing it to be applied uniformly across different declarative mapping scenarios.

B. Competency Questions and Requirements

The design of MMV was guided by a set of competency questions derived from observed metadata gaps in real-world mapping projects and evaluation campaigns. These questions capture the types of queries the ontology is intended to support and informed the identification of core concepts and relationships. Table 1 below summarizes the relationship

between these competency questions and the main ontology elements they motivated.

Table 1 Competency questions and Corresponding MMV elements.

CQ	Competency Question	Relevant MMV classes and properties
CQ1	What is the purpose and scope of a mapping project?	*mmv:MappingProject*, *mmv:hasAnalysis*, *mmv:hasRequirement*, *mmv:hasAssumption*
CQ2	Which domain and resources does the mapping operate on	*mmv:MappingProject*, *mmv:DataSet*, *prov:used*
CQ3	Which mapping artefacts were produced and in which format?	*mmv:MappingArtifact*, *mmv:mappingType*, *mmv:mappingLanguage*, *mmv:version*
CQ4	Which lifecycle activities led to the creation of the mapping?	*mmv:AnalysisActivity*, *mmv:DesignActivity*, *mmv:DevelopmentActivity*
CQ5	How was the mapping tested or validated	*mmv:TestingActivity*, *mmv:testsArtifact*
CQ6	How was the quality of the mapping assessed?	*mmv:EvaluationActivity*, *mmv:evaluatesArtifact*, *dqv:QualityMeasurement*
CQ7	Which tools were used during development and evaluation?	*mqv:MappingTool*, *prov:used*
CQ8	How has the mapping evolved over time?	*mmv:MaintenanceActivity*, *mmv:version*, *prov:wasDerivedFrom*
CQ9	Is the mapping suitable for reuse in a different context?	*mmv:MappingProject*, lifecycle activities, quality metadata

C. Core Concepts

At the centre of MMV is the concept of *MappingProject*, representing a coherent unit of work aimed at achieving a specific mapping objective. A *MappingProject* may involve one or more *MappingArtifact* instances, such as RML files, ontology alignments, or interlinking specifications.

MMV explicitly distinguishes between projects, activities, and artefacts. This modelling choice enables the association of multiple artefacts with shared context, including involved agents, tools, assumptions, temporal boundaries, and evaluation evidence.

D. Lifecycle Modeling

To represent the evolution of mapping projects, MMV models lifecycle phases as subclasses of *prov:Activity*. The ontology distinguishes between *AnalysisActivity*, *DesignActivity*, *DevelopmentActivity*, *TestingActivity*, and *MaintenanceActivity*, each capturing metadata relevant to that stage.

Early lifecycle activities document purpose, assumptions, requirements, and risks. Development activities capture implementation details such as mapping language and tools, while testing and evaluation activities record validation procedures and quality assessments. Maintenance activities represent updates and derivation relationships between mapping versions. Table 2 below summarises the lifecycle phases and the types of metadata captured.

Table 2 Lifecycle phases and metadata captured in MMV

Lifecycle Phase	MMV class	Main MMV Properties/Metadata
Analysis	*mmv:AnalysisActivity*	*mmv:hasRequirement*, *mmv:hasAssumption*, *mmv:hasRiskOrIssue*
Design	*mmv:DesignActivity*	Design decisions and reuse considerations (linked via *mmv:hasDesign*)
Development	*mmv:DevelopmentActivity*	*mmv:MappingArtifact*, *mmv:mappingType*, *mmv:mappingLanguage*, *mmv:version*, *mqv:MappingTool*
Testing/Evaluation	*mmv:TestingActivity* *mmv:EvaluationActivity*	Validation activities linked via *mmv:testsArtifact* *mmv:evaluatesArtifact*, *dqv:Metric*, *dqv:QualityMeasurement*
Maintenance	*mmv:MaintenanceActivity*	Updates and evolution actions linked via *mmv:hasMaintenance*

E. Reuse of Existing Vocabulary

MMV follows a reuse-oriented design by aligning its core concepts with established Semantic Web vocabularies. *MappingProject* and *MappingArtifact* are modelled as *prov:Entity*, while lifecycle phases are modelled as *prov:Activity*, enabling provenance-aware descriptions. Quality-related metadata reuse concepts from DQV[13], and MQV [14] is used to represent mapping tools and assessment activities. This alignment ensures interoperability while allowing MMV to introduce domain-specific structure.

F. Ontology Overview and Use Cases

MMV structures metadata around explicit relationships between MappingProject, lifecycle activities, and resulting MappingArtifact instances. A MappingProject is linked to its lifecycle phases via properties such as hasAnalysis, hasDesign, hasDevelopment, hasTesting, and hasMaintenance, enabling traceability between activities, artefacts, and evaluation outcomes (Figure 1).

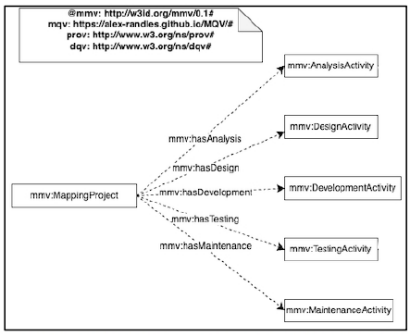


Figure 1summarizes the main classes and relationships in MMV.

Figure 1 presents a partial view of the MMV ontology diagram, generated using the CHOWLK [15] ontology diagramming tool. Due to space limitations, the complete diagram is provided in the online ontology documentation and specification[1].

To illustrate the application of MMV across different declarative mapping types, the following use cases show partial instantiations of the ontology. Due to space limitations, each example shows only a subset of the metadata rather than a full lifecycle description.

- Use Case 1: Ontology Alignment (Analysis Phase)

This use case (Figure 2) illustrates how MMV captures analysis-phase metadata for an ontology alignment task. A *MappingProject* is linked to an *AnalysisActivity* that records a textual description of the task, key requirements, and identified risks associated with aligning mouse and human anatomy ontologies. The example focuses on documenting early analytical considerations and constraints that inform later design and implementation decisions.

```
@prefix mmv:     <http://w3id.org/mmv/#> .
@prefix prov:    <http://www.w3.org/ns/prov#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix xsd:     <http://www.w3.org/2001/XMLSchema#> .

mmv:AlignProject a mmv:MappingProject ;
  dcterms:title "Anatomy Ontology Alignment" ;
  mmv:hasAnalysis mmv:AlignAnalysis ;
  mmv:projectStartTime "2025-06-01T00:00:00Z"^^xsd:dateTime .

mmv:AlignAnalysis a mmv:AnalysisActivity ;
  mmv:hasRequirement "Align mouse and human anatomy ontologies" ;
  mmv:hasAssumption "Overlapping anatomical scope" ;
  mmv:hasRiskOrIssue "Ambiguous labels across ontologies" .
```

Figure 2 MMV example for an ontology alignment use case

- Use Case 2: RML Uplift Mapping (Development Phase)

This use case (Figure 3) demonstrates how MMV represents development-phase metadata for an RML uplift mapping. A *MappingProject* is associated with a *DevelopmentActivity* describing the implementation of an uplift mapping from CSV data to RDF. The resulting *MappingArtifact* records the mapping type, mapping language, version, and reuse of an external mapping tool, illustrating how MMV documents how the mapping was implemented and which technologies were used.

```
@prefix mmv:  <http://w3id.org/mmv/#> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix mqv:  <https://alex-randles.github.io/MQV/#> .

mmv:RMLProject a mmv:MappingProject ;
  mmv:hasDevelopment mmv:RMLDevelopment .

mmv:RMLDevelopment a mmv:DevelopmentActivity ;
  prov:used mmv:RMLMapper ;
  mmv:hasRequirement "Generate RDF from CSV source" ;
  mmv:hasAssumption "Stable column identifiers" .

mmv:RMLMapping a mmv:MappingArtifact ;
  mmv:mappingType "uplift mapping" ;
  mmv:mappingLanguage "RML" ;
  mmv:version "1.0" .
```

Figure 3 MMV example for an RML uplift mapping use case

- Use Case 3: Interlinking (Testing/Evaluation phase)

This use case (Figure 4) illustrates how MMV captures testing and evaluation metadata for an interlinking task. A *MappingProject* is linked to a *TestingActivity* that validates a generated linkset, while an *EvaluationActivity* records quality assessment results using explicit quality measurements. The example demonstrates how MMV represents validation outcomes and quantitative quality evidence, supporting informed reuse and comparison of interlinking artefacts.

[1] https://sarahalzahranitcd.github.io/MMV1/

```
@prefix mmv:  <http://w3id.org/mmv/#> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix dqv:  <http://www.w3.org/ns/dqv#> .
@prefix xsd:  <http://www.w3.org/2001/XMLSchema#> .

mmv:LinkProject a mmv:MappingProject ;
  mmv:hasTesting mmv:LinkTest .

mmv:LinkTest a mmv:TestingActivity ;
  prov:used mmv:LinksetV1 .

mmv:LinksetV1 a mmv:MappingArtifact ;
  mmv:mappingType "interlinking" ;
  mmv:version "1.0" ;
  dqv:hasQualityMeasurement mmv:QM1 .

mmv:QM1 a dqv:QualityMeasurement ;
  dqv:computedOn mmv:LinksetV1 ;
  dqv:isMeasurementOf mmv:PrecisionMetric ;
  dqv:value "0.92"^^xsd:decimal .
```

Figure 4 MMV example for an Interlinking mapping use case

The complete MMV ontology, including its formal specification and documentation, is publicly available online[1] .

IV. Planned Expert Evaluation

A. Evaluation Goals

The planned expert evaluation aims to assess the conceptual clarity, coverage, and relevance of the Mapping Metadata Vocabulary (MMV) for documenting declarative mapping projects across their lifecycle. The evaluation focuses on whether MMV adequately represents metadata that domain experts consider important, including purpose, requirements, assumptions, validation, and maintenance information.

This study builds on an earlier survey of the Semantic Web community [5], which examined the relevance of different metadata elements for mapping-related tasks and informed the design of MMV. The current evaluation adopts a more focused perspective, examining whether MMV, as an ontology, captures these metadata requirements in a coherent and well-structured manner, while adhering to ontology engineering best practices such as conceptual clarity, appropriate abstraction, and reuse of existing vocabularies.

B. Evaluation Procedure

The expert evaluation is planned as a structured online survey targeting participants with experience in declarative mappings, ontology alignment, data integration, or knowledge graph construction. The primary objective of the evaluation is to assess the conceptual soundness, design quality, and coverage of the MMV ontology.

Participants will be asked to review selected parts of the MMV ontology documentation and respond to a survey combining three types of questions:

- Likert-scale questions assessing the clarity, relevance, and completeness of core ontology concepts and lifecycle phases.
- Targeted questions focusing on ontology modeling decisions, such as the level of abstraction and reuse of existing vocabularies.
- Open-ended questions allowing experts to identify missing concepts, ambiguities, or suggest improvements.

Domain experts from the Semantic Web community are invited to participate in the evaluation via an online survey available at the following expert Evaluation Survey[2] . The study has received ethical approval from Trinity College Dublin, and participation is voluntary and anonymous

V. Related work

From an ontology engineering perspective, a recurring challenge is how to represent processes and their outputs in a way that supports traceability, comparison, and reuse. Prior work has emphasized the explicit modelling of activities, artefacts, and agents when ontologies are intended to support analysis, governance, or reuse, rather than purely conceptual representation. Provenance models such as PROV-O provide a general foundation for describing such process-oriented information, but do not prescribe domain-specific lifecycle structures or relationships tailored to declarative mapping projects [16].

Large-scale evaluation initiatives and shared tasks constitute an important source of mapping artefacts, as illustrated by the Ontology Alignment Evaluation Initiative (OAEI). These initiatives generate large collections of mapping artefacts accompanied by narrative descriptions of experimental settings, evaluation procedures, and assumptions. While this supports result comparability, the lack of formal, machine-interpretable representations of project context and lifecycle information limits traceability and reuse of individual artefacts beyond their original evaluation setting [9].

Recent community-driven initiatives such as FAIR-IMPACT have explicitly addressed the need for improved documentation of mappings and crosswalks to support Findability, Interoperability, and Reuse [17]. FAIR-IMPACT provides guidelines and a practical framework for creating, documenting, and sharing mappings, including recommendations on metadata elements, identifiers, formats, and governance practices. However, these contributions are primarily methodological and do not define a formal ontology for representing mapping projects, lifecycle activities, and their interrelationships in a machine-interpretable manner[17].

MMV complements these efforts by providing a lifecycle-based ontology that models mapping projects as first-class entities and explicitly links activities, artefacts, and evaluation outcomes within a coherent structure. This positioning allows MMV to support consistent documentation, reasoning, and reuse of declarative mapping projects across different mapping types and evaluation settings, while remaining compatible with existing guidelines and provenance standards.

[2] https://forms.gle/Wg2KiZwgp84L6zBv8

VI. Toward a purpose-Centric Vocabulary

The development of MMV highlighted that explicitly capturing the purpose of an artefact is important not only for reuse decisions, but also for interpretation, evaluation, governance, and maintenance. In Semantic Web environments, artefacts such as mappings and datasets are frequently reused outside their original context. When purpose is not explicitly documented, this can lead to misinterpretation or inappropriate reuse.

While MMV captures purpose-related information within lifecycle activities, this work points to the value of modelling purpose as a first-class metadata element supported by a controlled vocabulary. A controlled vocabulary for purpose would promote consistency, comparability, and machine interpretability across artefacts, enabling clearer understanding of *why* an artefact was created and under which conditions it should be used.

As future work, we plan to explore a lightweight purpose vocabulary with predefined terms describing aspects such as project goal, application context, intended use, target audience, scope constraints, and reuse expectations. Such a vocabulary could complement lifecycle-based models like MMV and be applied more broadly across Semantic Web artefacts, beyond mapping projects alone.

VII. Conclusion and Future Work

This paper introduced MMV, a lifecycle-based ontology for representing metadata about declarative mapping projects. MMV models mapping projects as first-class entities linked to lifecycle activities and resulting artefacts, enabling structured documentation of context, assumptions, implementation decisions, and evaluation outcomes across the full mapping lifecycle. By reusing established Semantic Web vocabularies, MMV integrates provenance and quality metadata into a unified, extensible framework.

The ontology has been illustrated using two representative use cases ontology alignment and RML uplift mapping demonstrating its applicability across different mapping types and lifecycle phases. These examples highlight how MMV can support traceability and reuse assessment without being tied to a specific mapping language or toolchain.

Future work will proceed in two main directions. First, an expert-based evaluation will be conducted to assess the conceptual clarity, coverage, and relevance of MMV from an ontology engineering perspective. Second, a reuse-oriented evaluation will investigate whether MMV-based metadata improves users' ability to assess the suitability of existing mappings for reuse in new contexts.

More broadly, this work contributes to ongoing efforts to improve the transparency, comparability, and sustainability of declarative mapping practices by treating metadata as a core component of mapping projects rather than an afterthought.

Acknowledgment

The first author acknowledges financial support from Al Imam Mohammad Ibn Saud Islamic University and the Saudi government - represented by the Royal Embassy of Saudi Arabia - Cultural Bureau in Dublin. The second author is partially supported by the SFI ADAPT Research Centre (grant number 13/RC/2106_P2).

During the preparation of this work, the author used ChatGPT (GPT-4) and Grammarly for the purposes of grammar improvements. All AI-generated content was thoroughly reviewed, edited, and validated by the author, who takes full responsibility for the final manuscript and all its content.

References

[1] E. Iglesias, M.-E. Vidal, D. Collarana, and D. Chaves-Fraga, "Empowering the SDM-RDFizer tool for scaling up to complex knowledge graph creation pipelines1," *Semantic Web*, vol. 16, no. 2, p. SW-243580, 2025, doi: 10.3233/SW-243580.

[2] A. Dimou, T. De Nies, R. Verborgh, E. Mannens, P. Mechant, and R. Van de Walle, "Automated metadata generation for Linked Data generation and publishing workflows," in *LDOW2016*, CEUR-WS. org, 2016, pp. 1–10.

[3] F. Scharffe and D. Fensel, "Correspondence Patterns for Ontology Alignment," in *Knowledge Engineering: Practice and Patterns*, A. Gangemi and J. Euzenat, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 83–92.

[4] A. C. Junior, C. Debruyne, R. Brennan, and D. O'Sullivan, "FunUL: a method to incorporate functions into uplift mapping languages," in *Proceedings of the 18th International Conference on Information Integration and Web-Based Applications and Services*, in iiWAS '16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 267–275. doi: 10.1145/3011141.3011152.

[5] S. Alzahrani and D. O'Sullivan, "What metadata is needed for semantic and data mappings?," in *2023 IEEE 17th International Conference on Semantic Computing (ICSC)*, 2023, pp. 312–316. doi: 10.1109/ICSC56153.2023.00064.

[6] S. Alzahrani and D. O'Sullivan, "MetaMap: Facilitating Mapping Reuse and Lifecycle Documentation," in *2024 IEEE 18th International Conference on Semantic Computing (ICSC)*, 2024, pp. 323–328. doi: 10.1109/ICSC59802.2024.00058.

[7] S. Alzahrani and D. O'Sullivan, "MetaSEMAP: Usability Evaluation for Metadata Annotation Across the Mapping Lifecycle," in *First International Workshop on Users and Knowledge Graphs@ SEMANTiCS 2025*,

[8] S. Alzahrani and D. O'Sullivan, "Guiding LLM Generated Mappings with Lifecycle-Based Metadata: An Early Evaluation," in *Joint Proceedings of Posters, Demos, Workshops, and Tutorials of the 21st International Conference on Semantic Systems co-located with 21st International Conference on Semantic Systems (SEMANTiCS 2025), Vienna, Austria, September 3-5, 2025*, D. Chaves-Fraga, I. Heibi, D. Garijo, D. Collarana, A. A. Salatino, and S. Vahdati, Eds., in CEUR Workshop Proceedings, vol. 4064. CEUR-WS.org, 2025. [Online]. Available: https://ceur-ws.org/Vol-4064/kgnesy-paper4.pdf

[9] "Ontology Alignment Evaluation Initiative::Home." Accessed: Dec. 15, 2025. [Online]. Available: https://oaei.ontologymatching.org/

[10] "Ontology Alignment Evaluation Initiative::2025." Accessed: Dec. 15, 2025. [Online]. Available: https://oaei.ontologymatching.org/2025/

[11] "Ontology Alignment Evaluation Initiative::Anatomy." Accessed: Dec. 15, 2025. [Online]. Available: https://oaei.ontologymatching.org/2025/anatomy/index.html

[12] M. Poveda-Villalón, A. Fernández-Izquierdo, and R. García-Castro, "Linked Open Terms (LOT) Methodology (Version 1.0)." Ontology Engineering Group, Universidad Politécnica de Madrid, 2019. doi: 10.5281/zenodo.2539305.

[13] "Data on the Web Best Practices: Data Quality Vocabulary." Accessed: Dec. 16, 2025. [Online]. Available: https://www.w3.org/TR/vocab-dqv/

[14] "Mapping Quality Vocabulary." Accessed: Dec. 16, 2025. [Online]. Available: https://alex-randles.github.io/MQV/

[15] S. Chávez-Feria, R. García Castro, and M. Poveda-Villalón, "Chowlk: from UML-Based Ontology Conceptualizations to OWL," 2022, pp. 338–352. doi: 10.1007/978-3-031-06981-9_20.

[16] T. Lebo *et al.*, "PROV-O: The PROV Ontology," *W3C Recommendation*, 2013, [Online]. Available: https://www.w3.org/TR/prov-o/

[17] M. Martinková, P. Winstanley, C. Debruyne, and others, "D2.2 – FAIR Digital Objects for mappings between schemas and ontologies," FAIR-IMPACT, 2024. doi: 10.5281/zenodo.15111167.