Data, Data Processing, Data Provenance & Metadata.
On the digital representation of complex cultural data.

Abstract. The networking of objects facilitated by the Internet of Things as new as we might think. Every object that is catalogued for display within a cultural heritage institution is assigned entry-level data, along with further data layers on that object that each researcher will draw upon to create their research narratives, irrespective of their disciplinary background or bias. Within the community of researchers working with cultural data in particular, the desire to compare and aggregate diverse sources held together by a thin red thread of potential narrative cohesion, is only increasing. This poses challenges to information retrieval and contextualization in the digital age, it forces us to reassess the value and cost of metadata, and the consequences that accompany the use and reuse of digital data in a humanities or cultural research context. This paper discusses a number of the key barriers to the digital representation of complex cultural data.

Keywords: data — metadata — cultural computing — digital humanities —data processing — data provenance — cultural data — complex data — rich data — datafication

1. Introduction

The networking of objects facilitated by the Internet of Things isn't as new as we might think: the artifacts, natural and man-made, that inhabit the anthroposphere have long been connected by webs of associative, contextual data.[1] Take, for example, a simple seashell. Any given seashell can be said to have certain core properties about which there is likely to be broad consensus: it is hard, it is hollow, is has a certain color. But these core properties have the potential to take on—or be attributed with—very different meanings depending on who or what interacts with them, and what interpretation this individual might lay over these core properties. A child, for example, will most likely appropriate it for use in a manner that is very different from that of a fisherman. A craftsperson may take the item and, by inscribing designs on the shell, fundamentally alter the makeup of the item in and of itself, a factor that might later influence how a museum cataloguer would catalogue the item for display within a cultural heritage institution.

Put another way, each of the agents encountering the shell transforms it, appropriates it for their own needs or the needs of others they are themselves connected to, and in doing so each agent creates a narrative, capturing the meaning of the shell for them and for the moment in which they appropriate it. On this basis, narrative can here be understood as the story we tell about our data. As Jesse Rosenthal puts it

> narrative flaunts its human mediation. The term suggests communication—between a narrator and an implied narratee—and intention. More importantly, narrative declares itself as a retelling of something that had already existed in another form (2).

[1] This paper is a revised and extended version of a paper published in the MTSR2017 Conference: Edmond, J. and Nugent-Folan, G. (2017a) Data, Metadata, Narrative. Barriers to the Reuse of Cultural Sources. In Garoufallou, E., Virkus, S., Siatri, R., Koutsomiha, D. (Eds.). Metadata and Semantic Research. 11th International Conference, MTSR 2017, Tallinn, Estonia, November 28–December 1, 2017, Proceedings. Communications in Computer and Information Science (CCIS), Vol. 755, 253-260. Springer International Publishing, Cham, 2017.

These stories, and the associated layers of human mediation, however divergent they may be, share certain essentials that are specific to the object they relate to. There is a data layer—or data layers—that narratives will share and exploit to some extent, though perhaps not equally. To return to our example, a marine biologist will prioritise different fundamental characteristics (data) than a child collecting shells on a beach. In addition, in the course of the object's path through these narratives, it may be enhanced, altered, or otherwise transformed in ways that are driven by specific agents or external forces, but which may not be apparent to later finders: the decorative layer added by the designer may or may not appear to a later user as the result of human intervention.

This metaphor of the seashell with its layers of data, transformation, and narrative serves to highlight the forces that circulate around and shape our relationships with digital objects. Their interplay is not only useful for reimagining some of the challenges of information retrieval and contextualization in the digital age, but also for progressing an understanding of the value and cost of metadata and the use and reuse of digital data in a humanities or cultural research context.

2. Data: Slippery as a Fish?

A key part of the process of moving from phenomenon to data to narrative involves moving from an object or event to a document, that is, to the documentable data on the object or event. Suzanne Briet, in her groundbreaking work *Qu'est-ce que la documentation*? [*What is documentation?*] identifies this process when she makes her distinction between entity and document:

> Is a star a document? Is a pebble rolled by a torrent a document? Is a living animal a document? No. But the photographs and the catalogues of stars, the stones in a museum of mineralogy, and the animals that are catalogued and shown in a zoo, are documents (10).

All objects are not created equal: indeed, their variability is often a large part of their charm. In order to comply with computational systems, however, data must be equitable and interoperable in order to facilitate wider reuse and integration. The thing itself, once it has undergone a digital documentary process that essentializes it down to its manifestation(s) as data (a process we refer to as 'datafication'), is therefore less likely to contain reference to the individual, granular facets of the object that make it unique and that contribute to its richness as an item. How then do we capture the richness and idiosyncrasy of these items in a computation-driven environment and in digital systems that require sameness and systemization?

An agent encountering this datafied representation of an object or thing perceives and draws upon the data layer(s) they apprehend to create their own narratives. But each agent will likely perceive or interpret this data layer differently and, as a result, will think and speak about the object differently; accepting different core principles as input to their multiple narratives. While datafication allows us to draw from objects many of us are unlikely to otherwise have access to, be they records of past events, rare objects or complex cultural or scientific treatises, not everything can be datafied, and almost nothing can be datafied completely and absolutely. We are reliant on the process of datafication to make the item available *as data*, but this very process in and of itself is necessarily incomplete, as it is impossible to fully document or datafy the object. While population counts or environmental sensor readings may be immediately recognizable as data, comforting in their regularity and quantitative nature, they do so as the result of a documentation process designed to precisely to have this effect, that is, to accommodate partial representations rather than to fully represent the people or climate that has been thusly sampled; itself a near impossible task. How an object feels, the impact it has on a sensory level, together with any idiosyncratic facets of its provenance that have contributed towards shaping and reshaping it as an object will simply not be available any more.

This problem is further exacerbated by the fact that the term "data" is ubiquitous across an entire spectrum of research disciplines—from computational linguistics to physics and beyond—but is consistently interpreted differently, used indiscriminately in different contexts or different stages of the research process, or to refer to different things. Given the cross-disciplinary nature of data, this tendency for "data" as a term to be understood differently not just across disciplines, but *within* disciplines and even on a researcher by researcher basis, is hugely problematic to those looking to create truly inter-disciplinary digital research environments. What a computational linguist considers to be data (text, parallel corpora, machine translation output, or the annotations of professional translators, for example) will differ from what an anthropologist considers to be data, or what a historian considers to be data. This overdetermined network of data definitions engenders confusion and disorganization that inhibits inter- and cross-disciplinary dialogue, research, and collaboration. Furthermore, discussions of this very issue rely on discourse that is already polysemic to the point of blocking, rather than facilitating, understanding, can also be alienating to those approaching these debates from an information or computer science background; the very people charged with facilitating access to these materials.[2]

As Christine Borgman observes, more useful or proactive definitions of data are to be found in industry, with "The most concrete definitions of data [...] found in operational contexts" (20). However, such operational definitions of data are necessarily pragmatic and, for the most part, discipline specific, which means that the problems encountered on a discipline-specific small scale environment will be magnified and potentially misaligned when they are applied on an inter-disciplinary level. These operational definitions of data are also not definitions *per se*, but reflect discipline or industry-specific archival principles. Concordantly, these principles are often unclear in relation to what is and is not data, with data here being specific to the requirements of the specific industry or production process. Arguably this also delimits the re-interpretability of the data by presenting it in the context of an industry-specific database or dataset. For example, how a representative of the seafood industry speaks about seashells, and the type of data they would seek out on what are, for them, byproducts of food production, will differ drastically from the type of data an archivist working on these same seashells within a Natural History Museum would consider relevant for inclusion in a catalogue or database. Further still, because it is often the entity's placement within a database or other knowledge organization framework that causes it to be viewed as having the status of data, there is a high degree of contextual input involved in the demarcation of this material as data, with metadata taking on a prominent role in the assignation of data *as data*. In a digital environment, anything can *be data* once it is entered into a system *as data*. This highlights the central role and influence of metadata in a digital environment: metadata, as the data on data, points us towards the data, it designates data as data, revealing or obscuring aspects of the documentation, which has itself already revealed or obscured aspects of the original object.

3. Metadata and the Shifting Sands of Meaning

[2] See Edmond & Nugent-Folan, 2017b.

The previous section established that aside from the object itself, we have the entry-level data and the data layer or layers on that object that each agent will draw upon to create their narratives, irrespective of their disciplinary background, objectives, or bias. This is where the question of metadata comes in. The goal of metadata is to align and describe the data layers of objects so that users can find the material they need or the material that is relevant to their needs or objectives. Metadata supports, or should support, the formation of narratives by harnessing the process of transforming objects or experiences into documents (data). This data will be more homogenous than that which it documents, which, in theory, primes it to become more widely available. According to William Uricchio "data would be meaningless without an organizing scheme" (125). Metadata expresses such an "organizing scheme" to facilitate accessibility and discoverability of data, as well as its subsequent transformation into narratives. Indeed, by helping to demarcate data *as data*, and by organizing it within the specific context of a database or other knowledge environment, metadata has also arguably initiated the process of narrativisation, pointing the user toward data that may be considered similar or proximate within the database, while at the same time potentially directing the user away from (obscuring) other data within the same environment. To a certain extent, then, metadata influences the data it organizes and can thus be described as performative, having the potential not only to situate proto-data as data proper, but to indicate what it is about that data that one may find relevant, useful, or important. That it can also hide or conceal facets of data that are not as readily datafied or conveyed via the medium of metadata, should also be noted.

A rich information environment requires us to seek ways to reduce noise and enhance signal: this is what metadata does, as do other strategies that to focus on identifying and highlighting patterns within a text, such as data visualization, data mining, semantic uplift, or any number of others. But whose signal? Whose noise? Registries of digital objects created to mimic analogue finding aids are an essential part of the information retrieval landscape researchers face today. But they are also holdovers from a time when the affordances of the dominant technology were very different, and where a "natural curation" resulted from demographic and technological factors of the day. In the current context of largely unfettered access to digital data, this metadata can, as David Ribes and Steven Jackson observe about data, "become a sort of actor, shaping and reshaping the social worlds around them" (148) shaping and reshaping those fundamentals that underpin the range of possible narratives that can be created. If the interests of humanities researchers make every unit of data hold speculative value for the researcher, then we can no longer rely on "natural" curation via human limitations as we strive to make everything that is of value, be that value speculative or proven, potentially accessible within the environs of the database. After all, Raley notes that "Data cannot 'spoil' because it['s value] is now speculatively, rather than statistically, calculated" (124). It can, however, become hidden, or be rendered latent within an archive as a result of the very information architecture employed to facilitate its inclusion and findability within that archive.

Metadata standards have a mediating effect on the data that are accessible within the archive. They influence how we approach and conceptualize data. A failure to flag or fully account for data complexity leads to blinds spots within the archive. There is naturally a dual-threat here in the form of over- or under-describing: Over-describing risks losing the central signal from the material, privileging its uniqueness over its similarity to other cognate material, while under-describing naturally results in reduced findability. It is necessary then to interrogate metadata's capacity to both delimit and flag data complexity and concordantly, to identify pragmatic approaches that avoid delimiting data while endeavoring at all times to maintain the capacity for the data within a given system to display and communicate optimum semantic complexity. This is the sweet-spot that balances curation and complexity.

Not every person who approaches our seashell will be interested in the seashell in and of itself, just as not every agent that approaches an object within an archive or collection has the same motives, background, or preferences. For some then, the interest lies not in the data per se, but in how it's used: both in terms of the history of its usage (provenance), and its present-day functionality within the context of a database that has often been purpose built. These are the things metadata can either leave out or, in the case of the Shoah Foundation Visual History Archive (VHA), the metadata can be tailored to accommodate the interests of its primary audience. In "The Ethics of the Algorithm" Todd Presner provides detailed analyses of "the metadata scaffolding and data management system [...] that allows users to find and watch testimonies" (179) of Holocaust survivors. These include human-assigned "hierarchical vocabularies to facilitate searching at a more precise level" (176). Presner's concern is with the ethical implications of disassociating content and form: "Such a dissociation is not unique to the VHA but bespeaks a common practice in digital library systems and computation more generally, stretching back to Claude Shannon's theory of information as content neutral" (188).

Again however, this alternative metadata is tailored to a specific audience. It may be difficult, and extremely time consuming, to attempt to circumvent the narratives being suggested by the metadata. Even within this purposefully curated system designed to counteract "the impulse to quantify, modularize, distantiate, technify, and bureaucratize the subjective individuality of human experience" (179), it is still possible to hide materials that do not align with your usage-intention such as for example any "content that the indexer doesn't want to draw attention to (such as racist sentiments against Hispanics, for example, in one testimony)" (192). This again highlights the implications of "natural" curation via human limitations.

4. The Barnacled Shell: When Data Becomes Narrative, when Narrative becomes Data

The urge to create narrative from data is deeply set in human nature (Kahneman, 2013). Less recognized is what appears to be an equally innate ability to view the documentations that we encounter as somehow pristine, untouched by narratives: as "pure" data. In reality, the truth is often anything but: rather like the barnacles that colonize and give shape to a sea shell, the objects we study have complex histories, particularly when it comes to cultural data or objects that may have passed through hundreds of hands. These objects may have been subject to rebinding, restoration, damage, losses, or intentional edits, yet we still may think of the data we extract from them as somehow "raw," as though untouched by the hands or subjectivities of others. Brine and Poovey capture this paradox in their description of the so-called "data scrubbing" or "data cleaning" process as one "of elaboration and obfuscation" (73); it elaborates certain facets of the material, and obfuscates others. Borgman makes explicit that each and every decision made in the handling and rendering of data has consequences:

> Decisions about how to handle missing data, impute missing values, remove outliers, transform variables, and perform other common data cleaning and analysis steps may be minimally documented. These decisions have a profound impact on findings, interpretation, reuse, and replication (27).

In the sciences, the cleaning/ scrubbing of data is considered standard practice; so much so that it is often taken as a given and, as noted above, minimally documented.[3] While the knowledge that data is always already "cleaned" or "scrubbed" is implicit in disciplines such as economics, in the humanities, cleaning of data does not receive as much (or any) attention or acknowledgement, nor is it necessarily regarded positively. Contradictions and confusions are rampant across humanities disciplines not only with respect to what data is, how data becomes data, whether it needs to be cleaned or scrubbed, whether the fact that this scrubbing takes place is or is not implicit to the discipline, whether is an scientifically inappropriate manipulation, and whether original context, abstraction, or recontextualizing are integral functions for the treatment of data. How do we retain cognisance of the changes brought about by datafication and preparation, and of that which has been scrubbed away?

As Jennifer Edmond notes in "Will Historians Ever Have Big Data?" semantically and contextually complex cultural data are precisely the materials humanities researchers thrive on:

> How is this level of uncertainty, irregularity and richness to be captured and integrated, without hiding it "like with like" alongside archival runs with much less convoluted narratives of discovery? Who is to say what [...] is "signal" and what "noise"? Who can judge what critical pieces of information are still missing? (98)

These are the very materials whose provenance and idiosyncrasies as objects present us with problems in terms of how to incorporate their complexities into information systems, because the acts of datafication and the curation involved in creating a workable, searchable organization framework, invariably highlight certain facets of the material above others. Scholars in the sciences have acknowledged the issues surrounding the cleaning and processing that the digital documentation of objects undergo in order to be operable and inter-operable as data, but there appears to be a lack of concordant material addressing, acknowledging, and accounting for data processing in the humanities. In a field that is still reluctant to even adopt the term "data," the machinations and reshaping or re-contextualization of data remain under-acknowledged, and rarely explained or justified. What also needs to be addressed is whether the cleaning data undergoes is (or should be) reversible.

In terms of functioning data definitions in operation in the physical sciences at the moment, the NASA's Earth Observing System Data Information System (EOS DIS)[4] is perhaps one of the most thorough in terms of how it incorporates data processing into the system of classification: it not only acknowledges the levels of processing material undergoes to become data, but tiers this scrubbing or cleaning process, therein acknowledging that some material undergoes more extensive modification than others, and maintaining traceability to the source context or environ from which the so-called "native data" was extracted. The material recorded using the EOS DIS sees "Data with common origin [...] distinguished by how they are treated" (Borgman, 21). In other words, the data defined is by what researchers do to the material to make it data. In accordance with the EOS DIS, a researcher can opt for data at levels between 0 and 4, or even further back than the 0 phase, opting instead for native data (level pre-0 data?).

That said, this approach is not without its problems from a provenance perspective. Firstly, it is incomplete because of the presence of a level that precedes "level 0"; what is

[3] See Ribes & Jackson (2013) for further discussion of the cleaning that takes place in long-term data gathering projects.

[4] "NASA EOS DIS Data Processing Levels," https://science.nasa.gov/earth-science/earth-science-data/data-processing-levels-for-eosdis-data-products.

referred to in passing by Borgman as “native data,” a phrase that is both problematic (having as it does unpleasant connotations akin to those that accompany the use of the term “primitive” or “primitivism” in art) and acute. “Native data” retains emphasis on the importance of context apropos data, whether that context be “native” or in the form of the context(s) it acquires when it transitions from a native to a non-native environment: “Some scientists want level 0 data or possibly even more native data without the communication artefacts removed so they can do their own data cleaning” (22). However, while the distinctions between levels are relatively explicit, they only pertain to the onset of the research, the point where data is *gathered*. Thereafter the data is considered raw, until it is subjected to further processing:

> Although NASA makes explicit distinctions between raw and processed data for operational purposes, *raw* is a relative term, as others have noted [...] What is 'raw' depends on where the inquiry begins. To scientists combining level 4 data products from multiple NASA missions, those may be the raw data with which they start. At the other extreme is tracing the origins of data backward from the state when an instrument first detected a signal (26).

Interestingly, not one of the categories employed above has an analogous one in the humanities (aside from the rather loose concept of primary, secondary and tertiary sources), though that is not to say that a clear, lucid gradation of data that distinguishes how the material has been treated, or at least flags the fact that the data has been subjected to transformations, would not be beneficial for humanities researchers.

5. Conclusion and Recommendations

The challenges we present are not new: but the fact that they are acknowledged does not mean that they have no negative effect on our ability to access and reuse data, nor that we are moving steadily toward their resolution. Within the community of researchers working with cultural data, the desire to compare and aggregate diverse sources held together by a thin red thread of potential narrative cohesion is only increasing. The systems and standards promoted as enablers of this progress are generally imports from other epistemic cultures, however, and often are poor matches for the whole of the concerns cultural objects and data give rise to. It is integral that data, metadata, data cleaning and processing, together with the respective narratives these entities and activities both facilitate and engender, continue to function as enablers of research and that the barriers to meaning-making outlined in this article are minimized wherever and however possible.

In order to better facilitate digitally-delivered large-scale inter- or trans-disciplinary research infrastructures, there needs to be consensus in terms of what we speak about when we speak about data, irrespective of how difficult it is to “define” it when it comes to the diverse cultural resources that fuel humanities research. Furthermore, not everything of value for the study of human culture can or will be digitized. The digital record, the datafied material that makes up out databases, must therefore somehow incorporate that which is hidden to the digital eye. In its capacity to direct us towards material of interest, or to provide us with data on the data we encounter within a research database, metadata or its functional cousin, the registry of distinct or entities found in a given artifact, has the potential to play an integral role in ensuring that data that are not digitised or shared do not become “hidden” from aggregation systems. Metadata sees often ingenious attempts to facilitate findability within the context of a database, but it is necessarily even more selective than the data it points us towards: after all, fully representative metadata of a data entity would result in the metadata reproducing the data in its entirety: which defeats the purpose of metadata as an organizing schema that facilitates findability, particularly in big data environments. Lastly, there is a need for more transparency and accountability regarding data cleaning and the documentation of provenance, so as to ensure data with complex and idiosyncratic narratives are not overly homogenised and simplified. Balancing the need to maintain context and complexity with the efficiency of the suggestive outline may seem a paradox, or at least a challenge, but it will be a key underpinning for scholarship in the future that does not end up writing only the history of those on the winning side of the digital divide. Just as the sound of the sea is intrinsic to the curve of the seashell, we must also invent new shapes for the future of documentation that speaks both of itself and of its context, its origin, its journey to your hand.

6. Acknowledgements

This work has been funded by the European Commission as a part of the Knowledge Complexity (KPLEX) project, contract number 732340.

References

Borgman, C. (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. MIT Press.

Briet et al. (2006). *What Is Documentation?: English Translation of the Classic French Text*. Lanham, Md: Scarecrow Press.

Brine K. and Poovey, M. (2013) From Measuring Data to Quantifying Expectations: A Late Nineteenth-Century Effort to Marry Economic Theory and Data. In: Gitelman L (ed) "Raw Data" is an Oxymoron. MIT Press, pp 61-76.

Edmond, J. (2016). Will Historians Ever Have Big Data? In: *Computational History and Data-Driven Humanities. International Workshop on Computational History and Data-Driven Humanities*. Springer, Cham, 91-105. doi:10.1007/978-3-319-46224-0_9.

Edmond, J. and Nugent-Folan, G. (2017a) Data, Metadata, Narrative. Barriers to the Reuse of Cultural Sources. In Garoufallou, E., Virkus, S., Siatri, R., Koutsomiha, D. (Eds.). Metadata and Semantic Research. 11th International Conference, MTSR 2017, Tallinn, Estonia, November 28–December 1, 2017, Proceedings. Communications in Computer and Information Science (CCIS), Vol. 755, 253-260. Springer International Publishing, Cham, 2017.

Edmond, J. and Nugent Folan, G. (2017b) Digitising Cultural Complexity: Representing Rich Cultural Data in a Big Data environment. Paper presented at the *Ways of Being in a Digital Age — A Review Conference*, Oct 2017, Liverpool, United Kingdom. 2017. Available at: 〈 hal-01629459〉

Kahneman, D. (2013) Thinking Fast and Slow. Farrar, Straus & Giroux Inc.

Kirwan L (2013) Databases for quantitative history. In: Proceedings of the Third Conference on Digital Humanities in Luxembourg with a Special Focus on Reading Historical Sources in the Digital Age, Luxembourg, December 5-6, CEUR Workshop Proceedings, 1613.

Presner T (2015) The Ethics of the Algorithm: Close and Distant Listening to the Shoah Foundation Visual History Archive. In: Fogu C, Kansteiner W, Presner P (eds) Probing the Ethics of Holocaust Culture. Cambridge, Massachusets: Harvard University Press, pp. 175-202.

Raley R (2013) Dataveillance and Countervailance. Gitelman (ed) “Raw Data” is an Oxymoron. MIT Press, pp 121-145.

Ribes D, Jackson SJ (2013) Data Bite Man: The Work of Sustaining a Long-Term Study. In: Gitelman L (ed) “Raw Data” is an Oxymoron. MIT Press, pp 147-166.

Rosenberg D (2013) Data before the Fact. In: Gitelman L (ed) “Raw Data” is an Oxymoron. MIT Press, pp 15-40.

Rosenthal J (April 1 2017) Introduction: “Narrative against Data.” In Genre 50.1. Duke University Press, pp 1-18. doi:10.1215/00166928-3761312.

Uricchio W (2017) Data, Culture and the Ambivalence of Algorithms. In: Schäfer MT, van Es K (eds) The Datafied Society. Studying Culture through Data. Amsterdam University Press, pp 125-138.