Data, Metadata, Narrative.
Barriers to the Reuse of Cultural Sources.

Jennifer Edmond (0000-0001-9991-1637) & Georgina Nugent Folan (0000-0002-6216-9317)

Trinity College Dublin, Dublin, Ireland
edmond@tcd.ie ✉ & nugentfg@tcd.ie ✉

Abstract. The networking of objects facilitated by the Internet of Things isn't new. Every object that is catalogued for display within a GLAM institution is assigned entry-level data, along with further data layers on that object that each interactive agent (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 and presents the preliminary findings and recommendations of the EU Commission's Horizon 2020 funded KPLEX project (kplex-project.eu) in the field of knowledge complexity and cultural data.

Keywords: narrative · data · metadata · cultural computing · digital humanities

1. Introduction

The networking of objects facilitated by the Internet of Things isn't new: in fact, the artifacts, nature and man-made, that inhabit the anthroposphere have always been connected by webs of information. 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 they lay over these core properties. A child, for example, will appropriate it for use in a manner that is very different from that of a hermit crab. A fashion designer 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 would influence how a museum cataloguer, working decades later, might catalogue the item for display within a GLAM institution.

Put another way, each of the agents encountering the shell creates a **narrative**, capturing the meaning of the shell for them and for the moment in which they appropriate it. 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]

Roughly at the intersection of these narratives—these stories, or layers of human mediation—are, however, some shared essentials that all narratives draw upon. There is a **data layer** that narratives will share to some extent, though perhaps not equally, for example, a blind child won't know the color. 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, narrative, and transformation may seem an overly charged one, but the same forces that circulate around such evocative objects shape our relationships to 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?

An agent encountering an object or its representation perceives and draws upon the data layer they apprehend to create their own narratives. But any agent will likely perceive this data layer differently and, as a result, they retroactively think and speak about the data differently; identifying different "core principles" as input to their multiple narratives. A key facet of this problem is that the term data is ubiquitous, but is consistently interpreted differently, used in different contexts, or to refer to different things. Daniel Rosenberg [1] outlines the early history of "data" as a concept prior to the 20th century, and explores how it acquired its "pre-analytical, pre-factual status," while also clarifying that "facts are ontological, evidence is epistemological, data is rhetorical." Rosenthal [2] similarly presents data as an entity that "resists analysis." Christine Borgman [3] elaborates, stating that "Data are neither truth nor reality. They may be facts, sources of evidence, or principles of an argument that are used to assert truth or reality." Speaking in the context of data's relationship to fictional literary narratives, Rosenthal identifies data as a "fiction": "The fiction of data, the illusion of data." [2] In addition, Rita Raley [4] posits data as "performative" because "our data bodies [...] are repeatedly enacted as a consequence of search procedures. Data is in this respect performative." Anything can *be data* once it is entered into a system *as data*.

A problematic facet of the discourse on data is that the language used is often overly theoretical and alienating, and therefore not necessarily accessible to the practitioners looking to make data findable, accessible, reusable, and interoperable. Given the cross-disciplinary nature of "data" this is problematic and potentially alienating to those approaching these debates from an information or computer science background which may not encourage cross-disciplinary dialogue. More assured definitions, as Borgman [3] observes, are to be found in business: "The most concrete definitions of data are found in operational contexts." However, 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 on an inter-disciplinary level. How a representative of the seafood industry speaks about seashells will be different to how an archivist working within a Natural History Museum speaks about them. So these operational definitions of data are not definitions *per se*, but discipline or industry-specific archival principles. Concordantly, they are often unclear in relation to what is and is not data; and arguably they also delimit the re-interpretability of the data by presenting it in the context of a specific database or dataset. 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, with the metadata taking on a prominent role in the assignation of data *as data*.

If left unaddressed, the confusion and disorganization brought about by this overdetermined network of data definitions will have significant impact on future research programs and research infrastructures.[1] In order to better facilitate large-scale inter- or trans-disciplinary research infrastructures then, 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.

3. Metadata and the Shifting Sands of Meaning

To return to our shell analogy, aside from the object itself, we have the entry-level data, the data layer or layers on that object that each interactive agent will draw upon to create their narratives, irrespective of their disciplinary background or bias. Oftentimes, a necessary part of the process of moving from data to narrative involves moving from the **object** to a **document**.

[1] See [13] for a discussion on the impact of the absence of standardisation on the interoperability of historical trade data.

Suzanne Briet, in her groundbreaking work Qu'est-ce que la documentation? identifies the presence of this process when she makes her distinction between entity and document:

> Une étoile est-elle un document ? Un galet roulé par un torrent est-il un document ? Un animal vivant est-il un document ? Non. Mais sont des documents les photographies et les catalogues d'étoiles, les pierres d'un musée de minéralogie, les animaux catalogués et exposés dans un Zoo. [5]
>
> [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.] [6]

The thing itself may be found, by chance, and be essentialized down to its manifestation(s) as data, but only on an individualized, perhaps haphazard basis. Facilitating wider reuse and integration into digital systems requires systematization, but how do we capture the richness of these items in a computerized environment?

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 researchers can find the material they need or the material that is relevant to their research. Metadata supports, or should support, the formation of these narratives. According to William Uricchio [7] "data would be meaningless without an organizing scheme." Metadata is the "organizing scheme" that facilitates accessibility and discoverability of data, and its transformation into narratives. To a certain extent, metadata also influences the data it organizes and can thus be said to behave in a performative manner, having the potential to situate proto-data as data proper.

A rich information environment requires us to seek ways to reduce noise and enhance signal: this is what metadata does, but also data visualization, distant reading, semantic uplift, or any number of other strategies to focus on patterns within a text. 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 [8] observe about data, "become a sort of actor, shaping and reshaping the social worlds around them" shaping and reshaping those fundamentals that underpin the range of possible narratives that can be created. 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, 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 archive 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 has the same motives. For some then, the interest lies not in the data per se, but in how it's used. These are the things metadata can 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 [9] provides detailed analyses of "the meta-data scaffolding and data management system [...] that allows users to find and watch testimonies" of Holocaust survivors. These include human-assigned "hierarchical vocabularies to facilitate searching at a more precise level." [9] 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." [9] Again however, this alternative metadata is tailored to a specific audience, and even within this purposefully curated system designed to counteract "the impulse to quantify, modularize, distantiate, technify, and bureaucratize the subjective individuality of human experience," [9] 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)." [9] This again highlights the implications of "natural" curation via human limitations.

4. The Barnacled Shell: When Narrative Becomes Data

The urge to create narrative from data is deeply set in human nature. [10] Less recognized is what appears to be an equally innate ability to view what we find as somehow pristine, untouched by narratives: as data. In reality, the truth is often anything but: the shell we find takes its interesting shape and color from the barnacles that have colonized it, yet the data we find is still somehow "raw," as though untouched by the hands or subjectivities of others. Brine and Poovey [11] capture this paradox in their description of the so-called "data scrubbing" or "data cleaning" process as one "of elaboration and obfuscation"; 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. [3]

In the sciences 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.[2] 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. 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) and whether original context, abstraction, or recontextualizing are integral functions for the treatment of data. How do we retain cognisance of that which has been scrubbed away?

As Jennifer Edmond notes in "Will Historians Ever Have Big Data?" semantically and contextually complex cultural data is precisely the material humanities researchers thrive on; and again this presents us with a huge problem in terms of information systems management and curation:

> 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? [12]

If the interrogability of data makes everything potentially accessible, if the interests of humanities researchers makes every unit of data hold speculative value for the researcher, then we can no longer rely on "natural" curation via human limitations. After all, Raley [4] notes that "Data cannot 'spoil' because it['s value] is now speculatively, rather than statistically, calculated." 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.

Scholars have acknowledged the issues surrounding data cleaning and processing, particularly in the sciences, but there appears to be a lack of material addressing, acknowledging, and accounting

[2] See [8] for further discussion of the cleaning that takes place in long-term data gathering projects.

for data processing in the humanities where machinations and re-shapings or re-contextualization of data are under-acknowledged, rarely explained or justified, and often not reversible. What also needs to be addressed then is whether the cleaning data undergoes is (or should be) reversible in a manner skin to the material recorded using NASA's Earth Observing System Data Information System (EOS DIS),[3] where, as Borgman [3] notes "Data with common origin are distinguished by how they are treated." This sees data defined by what researchers do to the data 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?).

The EOS DIS is perhaps one of the most functional definitions of data available because 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 wherein the "native data" was extracted from. That said, this approach is not without its problems. Firstly, it is incomplete because of the presence of a level that precedes "level 0"; data that precedes "level 0" is 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." [3] 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. [3]

It remains to be seen as to how the EOS DIS table would look had it been compiled for the purpose of humanities research. Interestingly, not one of the categories employed 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, or that we are moving steadily toward their resolution. Certainly 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 KPLEX project (kplex-project.eu) is investigating these barriers to meaning-making. Our team has adopted a comparative, multidisciplinary, and multi-sectoral approach to this problem, focussing on key challenges to the knowledge creation capacity of cultural data such as the terms we use to speak about data in a cultural context, the manner in which data that are not digitised or shared become "hidden" from aggregation systems, the fact that data lacks the objectivity often ascribed to the term and the subtle ways in which data that are complex almost always become simplified before they can be aggregated.

---

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

Our initial results would suggest that the following measures contribute greatly to meaningful findability in cultural collections. First: we need to build upon existing momentum toward standardization, not necessarily of metadata approaches, but for the automatic uplift and storage to linked open data of information that might facilitate the serendipitous discovery of connections between otherwise dissimilar documents. Second: we must build upon existing provenance work for RDM such as the W3C standard for provenance (https://www.w3.org/TR/prov-overview/), or the activities of the RDA working group on Research Data Provenance. We must be more rigorous about documenting and sharing information about the transformations applied to data, so that we can access not just the data as it is now, but retrace its journey to its current state via a sort of "data passport." Third, we must work toward a state where cultural heritage institutional finding aids are able to converge with the secondary literature that discusses the collections represented there. The historic separation between scientific publishers and cultural heritage institutions is a huge barrier to an obvious opportunity to enhance the big data of the catalogues with the rich data of scholarly production. Finally, we must recognize that not everything of value for the study of human culture can or will be digitized. The digital record must somehow incorporate that which is hidden to the digital eye. This 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.

Acknowledgements. This work has been funded by the European Commission as a part of the Knowledge Complexity (KPLEX) project, contract number 732340. It bears an intellectual debt to Dr Michelle Doran of the KPLEX project team.

References

1. Rosenberg D (2013) Data before the Fact. In: Gitelman L (ed) "Raw Data" is an Oxymoron. MIT Press, pp 15-40.
2. Rosenthal J (April 1 2017) Introduction: "Narrative against Data." In Genre 50.1. Duke University Press, pp 1-18. doi:10.1215/00166928-3761312.
3. Borgman CL (2015) Big Data, Little Data, No Data: Scholarship in the Networked World. MIT Press.
4. Raley R (2013) Dataveillance and Countervailance. Gitelman (ed) "Raw Data" is an Oxymoron. MIT Press, pp 121-145.
5. Briet S (2014) Quoted in Day RE. Indexing It All: The Subject in the Age of Documentation, Information, and Data. MIT Press, 156.
6. Briet et al. (2006) What Is Documentation?: English Translation of the Classic French Text. Lanham, Md: Scarecrow Press, 10.
7. 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.
8. 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.
9. 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. Harvard University Press, Cambridge.
10. Kahneman D (2013) Thinking Fast and Slow. Farrar, Straus & Giroux Inc.
11. Brine KR and Poovey M, 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.
12. 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, 2016, pp 91-105. doi:10.1007/978-3-319-46224-0_9.
13. 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.