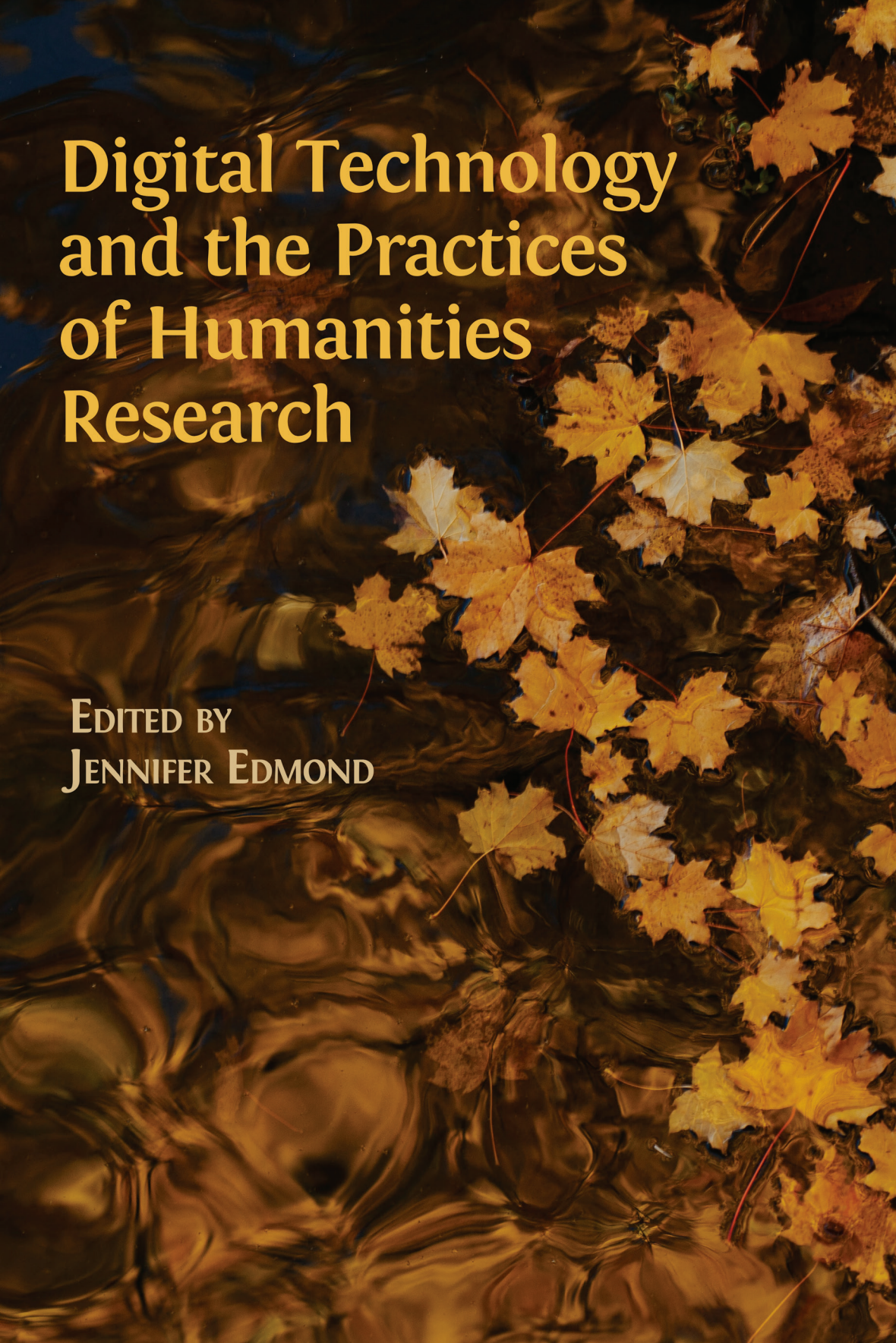
Digital Technology and the Practices of Humanities Research
Edited by
Jennifer Edmond

https://www.openbookpublishers.com

ISBN Paperback: 978-1-78374-839-6

ISBN Hardback: 978-1-78374-840-2

ISBN Digital (PDF): 978-1-78374-841-9

ISBN Digital ebook (epub): 978-1-78374-842-6

ISBN Digital ebook (mobi): 978-1-78374-843-3

ISBN Digital (XML): 978-1-78374-844-0

DOI: 10.11647/OBP.0192

Cover image: photo by Nanda Green on Unsplash https://unsplash.com/photos/BeVWHMXYwwo

Cover design: Anna Gatti

10. The Risk of Losing the *Thick Description*

Data Management Challenges Faced by the Arts and Humanities in the Evolving FAIR Data Ecosystem[1]

Erzsébet Tóth-Czifra

Realising the Promises of FAIR within Discipline-Specific Scholarly Practices

Since their inception in 2014, the FAIR principles (findability, accessibility, interoperability, and reusability) have come a long way in serving the global need for generic guidelines for data management and stewardship.[2] Addressing one of the grand challenges of scientific innovation, namely the need for infrastructure that supports the reuse of scholarly data, the FAIR principles have become increasingly influential since their formulation (created by a wide range of stakeholder groups who came together)[3] as a framework for the enhancement and optimisation of the digital ecosystem surrounding scholarly data publication.

1 I wish to thank Laurent Romary and Jennifer Edmond for their invaluable suggestions and comments on an earlier version of this manuscript.

2 Mark D. Wilkinson et al., 'The FAIR Guiding Principles for Scientific Data Management and Stewardship', *Scientific Data*, 3 (2016), https://doi.org/10.1038/sdata.2016.18

3 *Jointly Designing a Data FAIRPORT*, Workshop at Lorentz Center@Snellius, Leiden, 13–16 January 2014, https://www.lorentzcenter.nl/lc/web/2014/602/info.php3?wsid=602

 https://doi.org/10.11647/OBP.0192.10

The strong need for guidelines to enable and incentivise sustainable, connected, easily accessible, and cost-effective models of scholarly data curation was clearly reflected in the reception of the FAIR principles. The wide embrace and support for FAIR by governments, policy-makers, governing bodies, and funding bodies has not only made FAIR data, or 'FAIRification', a synonym for high-quality scientific data production, but has also fast-tracked the principles so they could make their way into global policies worldwide,[4] despite the many open questions their implementation leaves behind, and the palpable lack of agreed implementation plans and models at the level of different disciplines.

Considering how deeply they are embedded in the landscape of European scientific innovation and policy, the FAIR principles have the potential to make a substantial impact on the future landscape, as well as to shape the underlying dynamics of knowledge creation for the better. This chance, however, can easily be missed if the specific dynamics of scientific production in the humanities are not addressed in their discipline-level implementation.

With the goal of making FAIR meaningful, and helping it to realise its promises in an arts and humanities context, this paper describes some of the defining aspects underlying the domain-specific, epistemic processes that pose challenges to the FAIRification of knowledge creation in arts and humanities. In particular, by applying the FAIR principles to arts and humanities data curation workflows, it is demonstrated that, contrary to the principles' general scope and deliberately domain-independent nature, the principles have been implicitly designed according to underlying assumptions about how knowledge creation operates and communicates. In the following sections three such assumptions are addressed: first, that scholarly data or metadata is digital by nature;[5] second, that scholarly data is always

4 See, for example, European Commission, Directorate-General for Research & Innovation, *H2020 Programme Guidelines on FAIR Data Management in Horizon 2020* (26 July 2016), http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf; or Australian FAIR Access Working Group, *Policy Statement on FAIR Access to Australia's Research Outputs*, https://www.fair-access.net.au/fair-statement

5 See the 'Preamble' of the principles of: FORCE11, 'Guiding Principles for Findable, Accessible, Interoperable and Re-Usable Data Publishing Version B1.0', *FORCE11* (2014), https://www.force11.org/fairprinciples, where the eScience ecosystem is clearly indicated as being the domain of FAIR data management.

created and, therefore, owned by researchers;[6] and third, that there is wide community-level agreement on what can be considered to be scholarly data. The problems surrounding such assumptions in arts and humanities are the cornerstones for reconciling disciplinary traditions with FAIR data management. By addressing these assumptions one by one, this chapter contributes to a better understanding of the discipline-specific needs and challenges in data production, discovery, and reuse. These considerations may facilitate the inclusive and optimal implementation of high-level principles in a way that will serve to make the arts and humanities' disciplines flourish, rather than imposing limitations on their epistemic practices.

A Cultural Knowledge Iceberg, Submerged in an Analogue World

There is a fundamental difference between the epistemic cultures of STEM (science, technology, engineering, and mathematics) and those of the arts and humanities: namely, that in the arts and humanities the wide range of scholarly information artefacts, works of art, written documents of all sorts, recordings, annotations etc. — all of which can be broadly referred to as research data (in the sense used by Margaret E. Henderson)[7] — are not the autonomous products of research projects, but rather are deeply embedded in the cultural memory of Europe as well as the cultural and social practices of the institutions that preserve, curate, and (co)produce them. These institutions, commonly referred to as cultural heritage or GLAM (galleries, libraries, archives, museums) institutions — ranging from national libraries and archives down to small village museums or administrations — are typically not part of

6 Note that in the 'Preamble' there is no reference to data providers and data curators other than researchers (such as private or publicly funded providers of medical data, or curators of cultural heritage) nor are they mentioned among the stakeholders.

7 *Data Management: A Practical Guide for Librarians* (Lanham, MD: Rowman & Littlefield, 2016): 'Research data is data that is collected, observed, or created, for purposes of analysis to produce original research results' (p. 2). Other data definitions in a humanities context are more restrictive, for example, that of Christof Schöch (2013) in Christof Schöch, 'Big? Smart? Clean? Messy? Data in the Humanities', *Journal of Digital Humanities*, 2.3 (2013), http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/. As we will note later in this paper, the notion of research data is far from being straightforward in the arts and humanities.

the institutional landscape of academia. Despite this, the digital research ecosystem poses many challenges connected to the exploration and exploitation of the material and collections they hold; we do not need to get very far into the FAIR acronym to recognise these challenges.

The fact that these cultural sources and their enrichments are not merely representations of history, but also come with their own histories in terms of their creation and provenance, has serious implications regarding their visibility and shareability. Most importantly, the long tradition of cultural heritage data curation determines the way in which cultural resources are made available. According to a Europeana Foundation white paper from 2015, only ten percent of European cultural heritage is digitally available (300 million objects).[8] Therefore, the vast majority of cultural heritage data remain invisible on the digital horizon, which serves as the default domain of FAIR and scientific data management. Despite the combined digitisation efforts in Europe,[9] these numbers suggest that, for the foreseeable future, arts and humanities research will retain its hybrid nature, and encompass varying degrees of digital and analogue elements, thus calling for both automated and manual workflows and practices.

To give an example illustrating how much effort and investment is required to satisfy the basic requirements of data being digital in a cultural heritage context, Samuelle Carlson and Ben Anderson refer to two digitisation projects as cases in point: the CurationProject, which aimed at digitising and making available for study the records of a collection of more than 750,000 artefacts and 100,000 field photographs that had been collected since 1884; and the AnthroProject, where anthropological materials (including fieldwork notes, images, maps,

8 *Transforming the World with Culture: Next Steps on Increasing the Use of Digital Cultural Heritage in Research, Education, Tourism and the Creative Industries*, ed. by Beth Daley (The Hague: Europeana Foundation, 2015), https://pro.europeana.eu/files/Europeana_Professional/Publications/Europeana%20Presidencies%20White%20Paper.pdf. See also the same numbers in Fig. 3.6 in Gerhard Jan Nauta and Wietske van den Heuvel, *Survey Report on Digitisation in European Cultural Heritage Institutions 2015* (The Hague: DEN Foundation/Europeana/ENUMERATE, 2015), http://enumeratedataplatform.digibis.com/reports/survey-report-on-digitisation-in-european-cultural-heritage-institutions-2015/detail

9 European Commission, *Digitisation, Online Accessibility and Digital Preservation. Report on the Implementation of Commission Recommendation 2011/711/EU* (2013–2015), http://ec.europa.eu/information_society/newsroom/image/document/2016-43/2013-2015_progress_report_18528.pdf

and texts) from a range of countries were digitised and distributed through an online database and via DVDs.[10] In both projects, the major challenge was to build a well-structured, searchable database from their rather heterogeneous sources and records. This aim was realised as a rather long-term goal for both projects: the progressive digitisation, curation, and systematic documentation took thirty years in both cases.

Taking a step further towards findability, although digitisation is a preliminary first step in sharing knowledge, it alone does not guarantee the visibility and accessibility of cultural heritage data outside the walls of their hosting institutions. The aforementioned Europeana survey reveals that only one third (thirty-four percent) of digitised cultural heritage resources are currently available online, with barely three percent of these works suitable for real creative reuse; meaning, only this three percent has the chance to fulfil the discipline-specific measures of being FAIR.[11]

There are a number of cultural, social, legal, technical, and economic reasons that explain this small percentage of truly reusable cultural heritage data. These circumstances impact greatly on the working conditions of not only librarians, museologists, and archivists but also that of scholars who want to reuse and share data and content relevant to their research.

Legal Problems that Are Not Solely Legal Problems

The biggest obstacle in the productive reuse of digitised cultural heritage resources — from which many others derive — is the legal and ethical restrictions in which the usage conditions of cultural heritage sources are embedded. Determining the ownership status of research that is based on such material poses challenges in many cases. This is because the ownership status of research is, on some level, shared between the researcher who carries out the scientific analysis on the source materials, the institution that hosts and curates this material, and the people and cultures who give rise to the objects in question (e.g., photographers, and also the subjects of the photographs). Establishing

10 Samuelle Carlson and Ben Anderson, 'What *Are* Data? The Many Kinds of Data and their Implications for Data Re-Use', *Journal of Computer-Mediated Communication*, 12.2 (2007), 635–51, https://doi.org/10.1111/j.1083-6101.2007.00342.x

11 Daley, ed., *Transforming the World with Culture*, p. 9.

precise conditions for reuse on the basis of such a complex web of claims is, therefore, not an easy task.[12]

In addition to this complexity, provenance trails (i.e. a documented ownership and curation history of an artefact) are often embedded in historical practices, in particular in eras or contexts when the legal-ethical framework that defines present-day data exchange was either non-existent or irrelevant. Obviously, those handling these data could not know in advance that some information — for example, attribution or consent from the rights holders — needed to be collected: this requirement was only brought about by the digital age. Tracing back the provenance of such records is a time-consuming and difficult process filled with uncertainties and lack of clarity, especially in the case of collections inherited from other institutions.[13]

Furthermore, even in cases where the entity holding the legal right is clearly identifiable, given the great deal of legal uncertainty and variety present at the intersection of differing national legislations, and the changing landscape of intellectual property rights (IPR), in many cases researchers and curators are having difficulty 'translating' the legal statuses and license information of materials into research and publication workflows and terms of use. For instance, the legal statement 'In copyright, non-commercial use only' raises the question of where commercial use begins. Visual material under this legal status can certainly be integrated into PhD dissertations, but what about republishing such material on the researcher's website or in scholarly monographs?

The broad investigations of archival practices conducted within the framework of the Knowledge Complexity (KPLEX) project by Mike Priddy

12 To illustrate this complexity, let us cite here two examples from Carlson and Anderson's two aforementioned case studies: '[A researcher] has put a picture on the cover of a publication. He could be fined for that [by the community it originated from], because the artifact [*sic*] shows a ritual/secret process.'; and 'during her fieldwork in Malaysia, there was a photo collection (of a former local museum) that they wanted to sell to us. There were photos by tourists, army officers, etc. They think that they own every photo, but in our sense the photographer owns it, and we can therefore not show it' ('What *Are* Data?', 643).

13 This legal uncertainty in the identification of the legal statuses of cultural heritage material is clearly represented in the fact that in the Rights Statements framework, which has been designed specifically for cultural heritage data where the rights holder and the data provider are not always the same entities, four of the twelve standardised rights statements refer to unclear legal statuses. These are: 'In Copyright/Rights-holder(s) Unlocatable or Unidentifiable', 'Copyright Not Evaluated', 'Copyright Undetermined', and 'No Known Copyright'. See *Rights Statements for in Copyright Objects*, http://rightsstatements.org/en/

and Nicola Horsley reveal how such legal restrictions also affect technical and cultural aspects of data sharing in the cultural heritage domain.[14] In the context of developing support for interoperability frameworks via metadata standards and computational research methods, it is important to recognise that perceived or substantive legal barriers not only impact on the barriers for the reuse of content, but may prevent institutions from online metadata sharing as well. The identity of individuals or groups are often so deeply inscribed in the data that not even the highest level of abstraction can shield them. For example, some collection descriptions cannot be made available online because they contain biographical information about the person who donated them.

As the following excerpt from one of the interviews conducted in the KPLEX project indicates, such difficulties are either slowing down the standardisation procedure, increasing the manual curation effort required to produce sufficient and safe metadata, or simply preventing metadata sharing. This is especially problematic in the context of the FAIR recommendation that metadata should be open by default, even in cases of sensitive data.[15]

> [T]hese kinds of problems asked us to be able to make a choice between the collections, the metadata, which can be shared and the other ones and that took a lot of time. We weren't able to do that automatically, so these kinds of things, and it was totally impossible for us. So, for example, for [portal], to share metadata or to share documents with [portal]. It wasn't possible because of copyright issues or privacy issues.[16]

The need to fulfil legal requirements and to avoid the risk of penalties drives a conservative stance where there may be any uncertainty or grey area, and incentivises the practices of reduced sharing or holding data back out of a fear of lawsuits against, and legal liability of, the

14 Mike Priddy and Nicola Horsley, 'Deliverable D3.1 Report on Historical Data as Sources', *KPLEX* (2018), https://kplexproject.files.wordpress.com/2018/06/kplex_deliverable-d3-1.pdf. KPLEX is a Horizon 2020 project aimed at investigating ways in which a focus on 'big data' in ICT research elides important issues about the information environment we live in. The project focuses on four main themes: toward a new conceptualisation of data; hidden data and the historical record; data, knowledge organisation and epistemics; and culture and representations of system limitations.

15 Simon Hodson et al., 'Turning FAIR Data into Reality: Interim Report from the European Commission Expert Group on FAIR Data', *Zenodo* (2018), https://doi.org/10.5281/zenodo.1285272: 'The basic core is proposed as discovery metadata, persistent identifiers, and access to the data, or, at a minimum, metadata' (p. 57).

16 Priddy and Horsley, 'Deliverable D3.1 Report', p. 65.

respective institutions. The lack of a clear definition regarding the legal barriers puts a large portion of cultural heritage material into a minefield that neither practitioners in cultural heritage institutions nor scholars are willing to step into. The abandonment of certain research questions due to legal uncertainty, and the lack of accurate, transparent, and easily understandable conditions of access to documents, is an even bigger obstacle to FAIRification in the cultural heritage domain than the institution of legal protection that it aims to serve.

Case Study: The Removal of Photos from the CENDARI Project's Archival Research Guides due to a Lack of Information on their Reuse Conditions

The following case study from the CENDARI[17] project illustrates how legal, cultural, and data-management dimensions of non-transparency can lock away valuable and relevant cultural data so they cannot be reused, shared, and therefore sustainably preserved in the collective practices of heritage maintenance.

In February 2016, at the time of finalising the publication of CENDARI's Archival Research Guides,[18] scholars working on First World War materials were faced with a situation in which the ownership status of the illustrative images (found on the internet) was so unclear and inaccessible (even after detailed and repeated checks) that eventually the images in question had to be left out of the publication.

The online catalogues for the sources neither gave rights holder information, contact for publication permission, nor indicated the terms and conditions for the use of images.

This example illustrates the point that FAIR data is not necessarily open data, but data with clearly articulated reuse conditions. Notice that the problem here was not openness in the first place but a lack of transparency and proper data management that, in originating from external data providers, is out of the control of the researcher community. If the longevity of cultural heritage data is defined by their presence in scientific, cultural, and social discourses, then once we lose access to their reuse conditions, we lose them entirely.

17 *Cendari*, http://www.cendari.eu/

18 'Publicly Available Research Guides', *Cendari*, http://www.cendari.eu/thematic-research-guides/available-research-guides

The Risk of Losing the *Thick Description* upon the Remediation of Cultural Heritage

The advent of digital research infrastructures opened up a radically new frontier for the interactions with cultural heritage of both scholars and the public in an increasingly data-intensive and collaborative research ecosystem. As an active response to the impact of the digital age on scholarly and archival practice, a range of research data aggregation and discovery projects of different scopes and sizes have been created, such as: Europeana Collections,[19] IPERION CH,[20] and CENDARI.[21] They all have the mission to build bridges, interlinks, and networks (e.g., co-referencing systems, conceptual models, ontologies, semantic web frameworks) across different types of resources and institutions in order to enable the browsing of this heterogeneous content within a single search and discovery space. Although many of these infrastructures are facing sustainability challenges, their role in computationally-enhanced scholarly workflows is indispensable. Leveraging the power of big data and linked data approaches enables scholars to gain access to cultural heritage resources across institutional and national boundaries, and to explore new, macro-level perspectives and connections between distant events, communities, or traditions that could not have been made visible via traditional manual methods.

In addition to opening up new paradigms and epistemic models of knowledge creation, such research infrastructure initiatives also should be credited with having played a catalytic role in the development, promotion, and implementation of shared protocols and standards (like the Linked Open Data paradigm in arts and humanities)[22] to guarantee interoperability between heterogeneous data resources. Papers that report on data collection procedures for the research infrastructure projects EHRI (European Holocaust Research Infrastructure)[23] and

19 'Europeana Collections', *Europeana Collections*, https://www.europeana.eu/portal/?locale=en

20 'Iperion Homepage', *Iperion CH*, http://www.iperionch.eu/

21 *Cendari*, http://www.cendari.eu/

22 *Linked Data — Connect Distributed Data across the Web*, http://linkeddata.org/

23 Mike Bryant et al., 'The EHRI Project — Virtual Collections Revisited', in *Social Informatics*, ed. by Luca Maria Aiello and Daniel McFarland (Cham, Switzerland: Springer International Publishing, 2015), pp. 294–303, https://doi.org/10.1007/978-3-319-15168-7_37

CENDARI[24] provide an insight into the various challenges the participating projects and institutes had to face, as well as into the, sometimes, herculean efforts they made to put their records onto the world map of computationally remediated digital horizons.

Here, again, the standardisation of shared metadata has brought not only technical and financial challenges, but also epistemological challenges: the new ways in which cultural resources have been made available as a part of global networks affects the systems of discovery and knowledge creation. Following up on, and investigating the changing archival practices of cultural heritage institutions in the age of big data, the aforementioned KPLEX project[25] uncovered many important epistemological implications for the computational turn.

One of these has to do with losing control over the remediated records of archival knowledge and its complexity. In the course of traditional interactions, such as in-person visits or one-on-one consultations, archivists had the possibility of freely guiding the researcher through the collections and transferring all relevant knowledge to the specific research question. Since such mutual exchange-driven means of discovery are not possible in a computationally mediated context, researchers are left alone with the task of interpreting the specific datasets that had been harvested from institutions. Practitioners' concerns about misinterpretations and misuse of the data they had carefully curated were clearly and repeatedly indicated in the interviews.[26]

A speciality of data management in arts and humanities, therefore, is that it is highly dependent on external data providers, that is, the cultural heritage institutions.[27] As was also touched on in the CENDARI case study above, due to this dependence, certain aspects of data management and FAIRification efforts remain out of the control of researchers. In addition, the ways in which cultural heritage materials are made available to them define and, in many cases, impose limitations on the accessibility of complex knowledge structures. As a result of the separation of data from its context of creation (i.e. from the institution,

24 Jakub Beneš et al., *The CENDARI White Book of Archives* (2016), http://www.cendari.eu/sites/default/files/WhiteBook-Web.pdf

25 *KPLEX*, www.kplex-project.eu

26 Priddy and Horsley, 'Deliverable D3.1 Report' pp. 52–53, 64–68.

27 However, arts and humanities are not the only disciplines that are dependent on external data providers, see, for example, medical and health care studies.

its curators, and its wider provenance), collection descriptions that are part of the standardised and aggregated metadata remain the only reference points for the long history of records.

Creating descriptions is, therefore, a pivotal process, but also a complex task. Practitioners showed an awareness of how much the preparation of these online representations, and the alignment of the richest possible descriptions with their limited space and capacity, is an interpretative practice. As has also been pointed out by Wendy M. Duff and Verne Harris,[28] personal decisions made in the course of this knowledge transfer are inherently biased and will, therefore, foreground certain pieces of information, while leaving others sunk in analogue practices and tacit knowledge.[29] One thing, however, is clear: the separation of the data from the curators who bear this knowledge, instead providing an impoverished form of online access to such remediated knowledge representations, necessarily leads both to limitations in conveying their complexity and to simulacra that are misleading in their apparent completeness. This is crucial, because the loss of information is the loss of the continuous narratives of the origins and subsequent treatment of a source, which is critical to interpreting how it might be used in relation to other research sources — a central technique by which historical interpretations are corroborated and verified.

Consequently, the loss of this knowledge complexity imparts serious deficits in the reuse and interoperability potential of data made openly available by the hard work of curators, just as it may impoverish researchers' interpretation and understanding of the possible uses of sources. In other words, hiddenness and the loss of the *thick descriptions*[30] of holdings is a part of the process of making

28 Wendy M. Duff and Verne Harris, 'Stories and Names: Archival Description as Narrating Records and Constructing Meanings', *Archival Science*, 2.3 (2002), 263–85, https://doi.org/10.1007/BF02435625

29 This typically involves not only dynamics of foregrounding and backgrounding but also changes in scope and detail. 'Changing practice therefore carries risks of skimming over knowledge complexity to produce a simulacrum that represents less of an item's deviation from the collection in which it has been placed. In this way, differences between collections may become exaggerated as practitioners' "closeness" reinforces the unique value and identity of a collection as the smallest unit in their purview, while the complexity that distinguishes the unique value of items may be hidden.' Priddy and Horsley, 'Deliverable D3.1 Report', p. 83.

30 The term *thick description* is borrowed from cultural anthropology, a prominent subfield of the study of cultural heritage. The term was coined by the twentieth-century

historical and cultural records available for digital and computational discovery. Researchers in the arts and humanities always need multiple sources to verify interpretations, but this requires a deep knowledge of source provenance. Therefore, without complexity and context, the FAIR principles of maximum reusability and interoperability cannot be achieved on an epistemic level, even if they can be achieved technically.

As the results of the aforementioned Europeana survey suggest, the *thick description* of holdings is not the only layer of archival knowledge that might remain invisible or lost in a computationally mediated context of discovery. Practitioners' concerns about the non-digitised or offline substructure of an iceberg of knowledge, with the levels invisible below the water being forgotten and 'buried at deeper levels of accessibility during this transitional period' were clearly articulated in the KPLEX interviews.[31] It is a serious threat that a new generation of scholars might lose this awareness of materials and knowledge structures that have submerged beyond the digital horizon, resulting in a situation where one has to know what it is one cannot find. The main danger of this effect is that it may skew research towards what is easily available, easy to find, and, ideally, available freely online. This would generate a further enrichment and even greater visibility of this yet very small fraction of cultural heritage. Such asymmetry and distortion can cause potentially irreparable damage to our understanding of human culture. As Jennifer Edmond points out in her 2015 study, such distortion effects are also arising from the fact that, contrary to the essentially transnational nature of historical research, the digitisation of cultural heritage has largely been funded, and continues to be funded, along national lines, and not every country or institution has access to the same resources.[32] This results in substantial differences in the digital and online footprint

philosopher Gilbert Ryle (1900–1976), but it was the anthropologist Clifford Geertz who developed the concept into an ethnomethodological key notion with sufficient explanatory power, in his seminal work *The Interpretation of Cultures* (Clifford Geertz, *The Interpretation Of Cultures*, rev. ed. (New York: Basic Books, 2000), pp. 9–10). Geertz described the practice of *thick description* as a way of providing cultural context and meaning that people place on actions, words, things, etc. *Thick descriptions* provide enough context so that a person outside the culture can make meaning of the behaviour. Since then, the term and the methodology it represents has gained currency in the social sciences and beyond, and so today, *thick description* is used in a variety of fields of cultural study.

31 Priddy and Horsley, 'Deliverable D3.1 Report', p. 79.

32 Edmond, 'Tradition and Innovation', pp. 2–9.

of the various institutional holdings: wealthier institutions might have a stronger representation and, therefore, impact on historical research than those who have limited access to funding. This, in turn, 'risks creating perverse incentives for historians that bring to mind the tale of the drunk looking for his lost keys under the lamppost — not because that is where they were lost, but because that is where the light is'.[33]

Amid FAIRification efforts, as we develop our knowledge creation ecosystem to the next level — from a human-scaled to a machine-actionable one — the lessons that can be learned from these insights are crucial, and not only for researchers in the arts and humanities. Being attentive, along with maintaining an attitude of critical reflection regarding overall progress and limited or immature cases of openness, may help identify phenomena and situations where the principles enshrined in the first two letters of FAIR, 'findability' and 'accessibility', come into conflict with the last letter, 'reusability'. If we want to play it right in the computational research ecosystem, the ability to recognise and amend such contradictions is an essential skill for all researchers and in all research practices. Allowing knowledge icebergs and *thick descriptions* to remain invisible beyond the digital horizon would be an unreasonable price to pay for the sake of a paradigm shift. Being aware of them is a guarantee that we will not have to pay this price and can realise the promises of the innovative revolution to the full, thus enabling new forms of scholarly insight and communication.

The Scholarly Data Continuum

The previous sections highlighted that, in contrast to the hard sciences, the initial data in the arts and humanities is *collected*[34] rather than *generated*,[35]

33 Ibid., p. 4.

34 This distinction and its epistemological consequences are also articulated in Johanna Drucker's study on *capta* versus *data* where capta is 'taken' (the term *capta* stems from the Latin word for 'to take'), constructed, and is rooted in the co-dependent relation between the observer and the experience, while data represents observer-independent models of knowledge given as a natural representation of pre-existing fact. See Johanna Drucker, 'Humanities Approaches to Graphical Display', *Digital Humanities Quarterly*, 5.1 (2011), http://digitalhumanities.org/dhq/vol/5/1/000091/000091.html

35 Claudine Moulin et al., *Research Infrastructures in the Digital Humanities* (Strasbourg: European Science Foundation, 2011), p. 5, http://www.esf.org/fileadmin/user_upload/esf/RI_DigitalHumanities_B42_2011.pdf

and thus the digitisation of cultural heritage is an indispensable base for research in these disciplines. However, considering the highly intertwined systems of knowledge representation and knowledge creation[36] — a phenomenon that is commonly referred to in arts and humanities discourse as the illusion or oxymoron of raw data[37] — it is rather difficult to decouple this base of cultural data from the layers of analysis built upon them.

Embedded within the practices of making cultural heritage material digitally available, there is a series of decisions cultural heritage curators have to make: they range from decisions on what and what not to preserve, choosing classification systems and metadata schemas, determining the way in which texts and artefacts are photographed; to the ways in which text corpora are transcribed, encoded, or the OCR (optical character recognition) is corrected. All of these decisions impose a perspective, and thus an influence, on our perceptions of, and access to, data within a research environment. The creation of digital objects for arts and humanities research purposes is, therefore, not an innocent practice: it is not merely a prerequisite for digitally-enabled research, but is an important scholarly activity in itself. The initial layer of interpreting, preparing, and pre-processing cultural heritage data is, therefore, provided by the heritage institutions, a process that enables and gives access to other layers of analysis and knowledge creation resulting from scholarly activities.

Within the current practice, these different layers of analysis are separated by institutional silos and only in the rarest cases can they

36 See discussion on the 'fuzzy, implicitly highly networked data' in the humanities that questions the separability of the data areas of primary- and intermediate-data-results in Patrick Sahle and Simone Kronenwett, 'Jenseits der Daten: Überlegungen zu Datenzentren für die Geisteswissenschaften am Beispiel des Kölner "Data Center for the Humanities"', *LIBREAS. Library Ideas*, 23 (2013), https://libreas.eu/ausgabe23/09sahle/. Sahle and Kronenwett argue that by digitising the research process, the various types of research data merge into a continuum where narratives and knowledge creation practices are present from the initial data to the research output publications and keeping this continuum together poses special challenges in data management and hosting infrastructure. The challenges in keeping together different mediums of knowledge creation, data and software in the first place is a general and major challenge in sustainable in reproducible data management and is a topic that deserves more detailed discussion than it can receive within the framework of the present paper.

37 Virginia Jackson and Lisa Gitelman, 'Introduction', 'in *Raw Data' Is an Oxymoron*, ed. by Lisa Gitelman, Geoffrey C. Bowker, and Paul N. Edwards (Cambridge, MA: MIT Press, 2013), pp. 1–14, https://doi.org/10.7551/mitpress/9302.003.0002

stay connected with each other. As a result, the actual continuum of the knowledge creation procedures of the cultural heritage domain is barely reflected in its infrastructure and data management practices.

A key recommendation in the FAIR principles, which aims to facilitate access to research data, is that data should be stored in trusted and sustainable digital repositories.[38] Taking the view from the researchers' side of cultural heritage knowledge creation, the landscape of outputs and throughputs show a rather fragmented picture. At the time of writing, the reference repository catalogue re3data.org lists 206 data repositories under the subject label 'humanities'; a relatively small number, not only in comparison with umbrella disciplines with more robust traditions of 'data-drivenness' such as life sciences (1,132 results), but also compared to the sibling disciplinary group, social and behavioural sciences (331 results).[39] The low number of repositories suggests lower demand for data sharing services, or, at least, a less established data sharing culture in the arts and humanities than in other fields of study.[40] On the other hand, however, several recent studies herald an increasing interest in data sharing in the arts and humanities at a global disciplinary scale.[41] For instance, in Ruth Mostern and Marieka Arksey's 2016 study,[42] which surveyed the target users of the Collaborative for Historical Information and Analysis

38 Hodson et al., 'Turning Fair DATA into Reality', p. 18.

39 *Re3data Registry of Research Data Repositories*, www.re3data.org

40 In their 2013 study investigating disciplinary differences in data management practices, Katherine G. Akers and Jennifer Doty arrive at similar conclusion. They found that in their university (Emory University) arts and humanities researchers tend not to store their data using university-based servers but instead rely heavily on computer/external hard drives and internet-based storage. Katherine G. Akers and Jennifer Doty, 'Disciplinary Differences in Faculty Research Data Management Practices and Perspectives', *International Journal of Digital Curation*, 8.2 (2013), 5–26 (p. 9), https://doi.org/10.2218/ijdc.v8i2.263

41 Rinke Hoekstra, Paul Groth, and Marat Charlaganov, 'Linkitup: Semantic Publishing of Research Data', in *Semantic Web Evaluation Challenge*, ed. by Valentina Presutti et al., Communications in Computer and Information Science (Cham, Switzerland: Springer International Publishing, 2014), pp. 95–100, https://doi.org/10.1007/978-3-319-12024-9_12; Sandra Collins et al., *Going Digital: Creating Change in the Humanities* (Berlin: ALLEA E-Humanities Working Group Report, 2015), p. 6, https://hal.inria.fr/hal-01154796

42 Ruth Mostern and Marieka Arksey, 'Don't Just Build It, They Probably Won't Come: Data Sharing and the Social Life of Data in the Historical Quantitative Social Sciences', *International Journal of Humanities and Arts Computing*, 10.2 (2016), 205–24, https://doi.org/10.3366/ijhac.2016.0170

(CHIA) database, ninety-four percent of the respondents indicated that they would consider putting their data in a repository.[43]

Understanding this large gap between intentions, real willingness, and practice is a key step towards the development of research data management services and recommendations that match humanities researchers' needs.

Data in Arts and Humanities — Still a Dirty Word?

Sharing data necessarily implies having or owning data. In addition to the aforementioned complexities in the shared ownership of primary sources, which forms a major hindrance to data sharing, having data or working with data is not always a straightforward process, especially in the traditional fields of arts and humanities. Iterated and large-scale surveys would be beneficial for assessing whether, and to what extent, the term 'data' is still a dirty word in the increasingly digital humanities disciplines and how the evolving landscape of open data and FAIR data policies impact and transform such conceptions of data.[44]

Surveys from the past five years[45] reveal a great deal of uncertainty in the arts and humanities researchers' conception of data and its

43 This seems significant progress over, for example, Diane Harley et al., *Assessing the Future Landscape of Scholarly Communication: An Exploration of Faculty Values and Needs in Seven Disciplines* (Berkeley, CA: Center for Studies in Higher Education, 2010), https://escholarship.org/uc/cshe_fsc. In this study, evidence is shown that historians are cautious about sharing work publicly until it is well-polished. Similar to many other fields in the arts and humanities, drafts are generally circulated by email among a small network of trusted colleagues for comment, feedback, and improvement. The study also points out how sharing habits are dependent on career stages; while graduate students and pre-tenure scholars may harbour fears that openly shared, in-progress work could be heavily criticised or poached, tenured scholars tend to be more comfortable with sharing early research ideas and other in-progress work. As concerns data sharing, the study argues that 'While scholars have varied opinions regarding the sharing of primary archival data, few scholars share their research notes, databases, or other intermediary interpretations of archival material; those who do usually wait until they have formally published their research' (p. 451).

44 Alicia Hofelich Mohr et al., 'When Data is a Dirty Word: A Survey to Understand Data Management Needs Across Diverse Research Disciplines', *Bulletin of the Association for Information Science and Technology*, 42.1 (2015), 51–53, https://onlinelibrary.wiley.com/doi/full/10.1002/bul2.2015.1720420114

45 Akers and Doty, 'Disciplinary Differences'; Mohr et al., 'When Data is a Dirty Word'; Hélène Prost, Cécile Malleret, and Joachim Schöpfel, 'Hidden Treasures: Opening Data in PhD Dissertations in Social Sciences and Humanities', *Journal of*

applicability to their own work.[46] Concerns and difficulties around the concept of data were clearly reflected in responses to the survey conducted by Jennifer L. Thoegersen in 2018 and published under the title '"Yeah, I Guess that's Data": Data Practices and Conceptions among Humanities Faculty'.[47] Here, humanities faculty members from the University of Nebraska-Lincoln were interviewed about their data management practices; all the participants expressed some level of uncertainty while talking about their own data management practices. For example, someone asked, 'Does that sound right?',[48] after providing a definition of data.

The study does not specify any information about the research practices of the faculty members, so the intriguing question is left open as to whether there is any correlation between data awareness and the level of integration of computational methods into the respective research workflows. Another relevant feature of arts and humanities research that may explain confusion around the notion of data is the great variety in the types of sources and information throughputs and outputs (laser scanner data, musical notations, voice recordings, annotations, critical editions etc.) produced by the wide ranging disciplines that come under the umbrella term of arts and humanities, as well as under the umbrella term data in computational research contexts.

The Critical Mass Challenge and the Social Life of Data

The intensifying discourse around data conceptions and data characteristics clearly indicates a shift in the paradigm towards data-driven and computational methods across the whole disciplinary range of the arts and humanities. Yet, there are still plenty of interrelated

Librarianship and Scholarly Communication, 3.2 (2015), http://doi.org/10.7710/2162-3309.1230; Jennifer L. Thoegersen, '"Yeah, I Guess that's Data": Data Practices and Conceptions among Humanities Faculty', *Libraries and the Academy*, 18.3 (2018), 491–504.

46 As Jennifer L. Thoegersen remarks, researchers in arts and humanities may not be comfortable describing their scholarly and academic work as data. A potential reason behind this is that in their data conceptions are tied to the prototypical data representations such as numerical or quantitative description of data. Thoegersen, '"Yeah, I Guess that's Data"', 492.

47 Thoegersen, '"Yeah, I Guess that's Data"'.

48 Ibid., p. 501.

issues that prevent data sharing in subject repositories (which are, as we have seen, central data services in the implementation of the FAIR principles) and hamper reuse in becoming an entrenched and integral part of scholarly practices. In their 2016 paper 'Don't Just Build It, They Probably Won't Come: Data Sharing and the Social Life of Data in the Historical Quantitative Social Sciences', Mostern and Arksey capture many such interrelated problems that define the current repository landscape in the arts and humanities,[49] which lingers in a vicious cycle of data repository failure. They make these observations in the context of quantitative historical research, but it is not a stretch to extend these insights to the multitude of scholarly communities in the arts and humanities, keeping in mind that they are not equally plagued with the problems described.

As has been pointed out in several other discipline-specific data management studies, there is a lack of incentives and rewards to dedicate to the considerable amount of time, effort, and expertise needed to prepare data for computational analysis and make it compliant with the standards and data models of the repositories.[50] Consequently, only a small user community is open to taking steps in sharing data and thus contributing to the development of repositories. As a result, the limited number of contributions coming from this small user base will not attract further communities to visit or contribute to them.[51] In addition,

49 Mostern and Arksey, 'Don't Just Build It'.

50 Robin Rice and Jeff Haywood, 'Research Data Management Initiatives at University of Edinburgh', *International Journal of Digital Curation*, 6.2 (2011), 232–44, https://doi.org/10.2218/ijdc.v6i2.199; Alex H. Poole, 'Now is the Future Now? The Urgency of Digital Curation in the Digital Humanities', *Digital Humanities Quarterly*, 7.2 (2013), http://www.digitalhumanities.org/dhq/vol/7/2/000163/000163.html; Catherine Anne Woeber, 'Towards Best Practice in Research Data Management in the Humanities' (unpublished master's dissertation, School of Information Management, Victoria University of Wellington, 2017), http://researcharchive.vuw.ac.nz/handle/10063/6620

51 Note that guaranteeing the presence of a target audience by reaching a critical mass of content was the recipe for success of the two academic sharing and networking platforms ResearchGate and Academia.edu, which even today are commonly used. We can learn a lot from the failures that underlie their conceptual design and what became visible only after they reached a critical level of user engagement. Although the original aim of both platforms was to help researchers go beyond paywalls and increase the availability of their research, the low entry thresholds (direct upload of PDFs, no custom metadata, no licensing options) conserved bad sharing behaviours (low awareness of copyright, which article versions are allowed to be legally shared, low awareness of the importance of licensing issues, support for freemium business models based on selling data on user behaviours) on such

repository developers and standardisation bodies then do not receive a significant enough input foundation from diverse sources that could serve as a sufficient and informative basis for developing infrastructural components (widely accepted metadata standards tailored to specific data types, for example, or analytical tools for opening up the boxes of deposited datasets etc.) such as could truly increase the visibility and discoverability of deposited data, and that could also connect them with other databases or datasets. This lack of momentum preserves the scattered landscape of subject repositories, and also maintains the status of repository users as an invisible or only slightly visible part of the wider disciplinary communities. This prevents their work and approaches from being both accessible and strongly represented to students and peers. In turn, it does not encourage them to share their data; thus, ultimately, the strongest appeal for the use of repositories is not able to work its charm.

Having been inspired by the 2003 study by Jeremy P. Birnholtz and Matthew J. Bietz,[52] Mostern and Arksey describe this complex phenomenon as the lack of the social life of data. Recognising the importance of having a community aspect around robust data sharing culture (wherein documents and deposited datasets are not only a means for delivering information, but are also meant for maintaining social groups and the professional exchange around them), they came to the important conclusion that repositories can only succeed as long as scholarly communities create social communities around them.[53]

---

a massive scale that it seriously slowed down the development and large-scale uptake of more sustainable, transparent, and legal ways of self-archiving (such as the use of preprint servers). For more discussion on such controversies see: Jonathan P. Tennant, 'ResearchGate, Academia.Edu, and Bigger Problems with Scholarly Publishing', *Green Tea and Velociraptors* (2 February 2017), http://fossilsandshit.com/researchgate-academia-edu-and-bigger-problems-with-scholarly-publishing/

52 Jeremy P. Birnholtz and Matthew J. Bietz, 'Data at Work: Supporting Sharing in Science and Engineering', in *Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work*, GROUP '03, ed. Kjeld Schmidt, Mark Pendergast, Marilyn Tremaine and Carla Simone (New York: ACM, 2003), pp. 339–348, https://doi.org/10.1145/958160.958215

53 These observations show congruency with the main findings of a much earlier study on the uptake and use of digital resources in the arts and humanities, namely the LARIAH project (Log analysis of Internet Resources in the Arts and Humanities; see a project description in C. Warwick et al., 'Evaluating Digital Humanities Resources: The LAIRAH Project Checklist and the Internet Shakespeare Editions Project', in *Openness in Digital Publishing: Awareness, Discovery, and Access. Proceedings of the 11th International Conference on Electronic Publishing*, Vienna, 13–15

This primarily includes peer evaluation of the deposited datasets. Data peer review is not only a vital step towards the acknowledgement and recognition of research data sharing, but, as their survey shows, it is also important in building user confidence, as seventy percent of historians responding to their survey indicated that a peer review process or citation option as part of the data submission process would increase their incentive to do so.

The idea of providing infrastructural support to bring the scholarly practices of data depositing and data peer review into closer proximity is also expressed in a checklist of recommendations in the Log Analysis of Internet Resources in the Arts and Humanities (LAIRAH) project. According to these recommendations, the ideal digital resource should be as follows:

1. it should have access to good technical support, ideally from a centre of excellence in digital humanities;
2. it should recruit staff who have both subject expertise and knowledge of digital humanities techniques; and
3. it should also retain this expert staff by having constant access to funds.[54]

Data peer review along these lines — that is, focusing on the support and joint development of transparent and good quality data creation without the power dynamics and the gatekeeping function that are causing serious challenges in the institution of the traditional article and book peer review[55] — could also be interpreted as a significant

June 2007, ed. by Leslie Chan and Bob Martens (Vienna, Austria: ELPUB, 2007), pp. 297–306, https://publik.tuwien.ac.at/files/pub-ar_7877.pdf). The project was based at UCL's School of Library Archive and Information Studies and was aimed at identifying the various factors (under the categories of content, user, maintenance and dissemination) that influence the long-term sustainability and use of digital resources in the humanities. Reaching a critical mass and gaining prestige within a university were found to be vital in the sustainability and longevity of digital infrastructures. In addition, the importance of good project staff and the availability of technical support have also been pointed out. As a result of the research, Warwick et al. ('Evaluating Digital Humanities') provided a checklist of recommendations to facilitate both the successful design of digital infrastructures and the recognition and culture around them.

54 Warwick et al., 'Evaluating Digital Humanities', pp. 302–03.

55 See, for example, Jonathan P. Tennant, 'The State of the Art in Peer Review', *FEMS Microbiology Letters*, 365.19 (2018), https://doi.org/10.1093/femsle/fny204

contribution to a more sustainable and more inclusive culture of research evaluation in general. At the same time however, the third LAIRAH recommendation stated above also indicates the serious sustainability challenges for such models in terms of funding. The ability to maintain, in repositories, both a technically and disciplinarily highly skilled expert staff, who have the capacity to provide a thorough evaluation of the massive number of data deposits that can be expected as a result of FAIR policies, does not seem to be a viable option. As a potential alternative, institutional data stewards[56] and data centres like the Leiden University Centre for Digital Humanities (LUCDH)[57] could at least partially fulfil this role.

An additional challenge in facilitating the culture of data evaluation in the arts and humanities, as has been pointed out by others, is that the scholarly practice of data peer review is still substantially lagging behind the traditional paradigm of research article publishing, which serves as academia's highest value currency.[58] Bringing these two forms and practices of scholarly communication, data sharing, and article or book publishing, closer to each other is a key step towards a more open, more connected, more transparent, and more sustainable research data management ecosystem.

The Risk of Losing the *Thick Description* — Again

Relying on domain-relevant community standards is critical to avoid having deposited datasets being buried in isolated 'data tombs', and to

56 Rec. 13 of the FAIR Data Action Plan (Hodson et al., 'Turning FAIR Data into Reality', p. 73.) recommends developing two cohorts of professionals to support FAIR data: data scientists embedded in those research projects that need them, and data stewards who will ensure the management and curation of FAIR data.

57 Researchers who need help or have questions regarding the critical use of digital technology and computational approaches in disciplines of the humanities can get support from the Leiden University Centre for Digital Humanities (LUCDH). A case study published in a recent collection of FAIR data advanced use cases from the Netherlands gives an insight into how this type of institutional support might work in an arts and humanities context. Melanie Imming, 'FAIR Data Advanced Use Cases: From Principles to Practice in the Netherlands', *Zenodo* (2018), 33–35, https://doi.org/10.5281/zenodo.1246815

58 E.g., Anne Baillot, 'A Certification Model for Digital Scholarly Editions: Towards Peer Review-Based Data Journals in the Humanities', *HAL* (2016), halshs-01392880, https://halshs.archives-ouvertes.fr/halshs-01392880/document

increase the social life of data by making it interoperable and connectible with other data sources. Achieving compliance with metadata standards is a prerequisite for improving the visibility, accessibility, interoperability, and linking of digital resources. Shared standards open up datasets for integration with research across different sectors, provide additional layers of context, and enable research methods that have not been previously available to the humanities.

However, aligning the application and use of repository standards with the long history of data curation cannot always be achieved without making compromises. In some cases, enforcing a commitment to shared standards can lead to a similar loss of detail and information, as was seen in the context of the aggregation of standardised and machine-interoperable metadata from cultural heritage institutions. In their 2014 and 2016 studies, Rinke Hoekstra and his co-authors investigated data sharing practices in the humanities and their compliance with linked discovery context.[59] They identify two cases in which the risk of losing provenance information is especially high.

First, when data is deposited in bigger, discipline-specific data curation projects with top-down standards (such as the North-Atlantic Population Project (NAPP), the Clio Infra repository, or the Mosaic project), Hoekstra et al. point out that the sheer scale of such databases and the top-down fashion of their data curation standards are not always suitable for smaller datasets created by individual researchers. This makes it difficult for them to share their research in a sustainable way.[60]

Second, not every researcher has equal access to the computational resources, expertise, and skills necessary to create and operate a digital data collection. To address this problem a number of low-barrier-to-entry repository data services have been created (e.g., EASY, Dryad, Dataverse, and Figshare). These services are important pillars of scientific data sharing infrastructure as they help to satisfy the growing demand for sustainable data sharing and archiving services. They enable easy data upload in most formats; ensure data is citable via

59 Hoekstra, Groth, and Charlaganov, 'Linkitup'; Rinke Hoekstra et al., 'An Ecosystem for Linked Humanities Data', in *The Semantic Web*, ed. by Harald Sack et al., Lecture Notes in Computer Science (Cham, Switzerland: Springer International Publishing, 2016), pp. 425–40.

60 Hoekstra et al., 'An Ecosystem', p. 426.

persistent identifiers, and also guarantee long-term archival storage. On the other hand, as argued in the earlier study, these generic-scope data sharing platforms bear hidden limitations on discoverability and productive reuse.[61] The first limitation is the result of the rather isolated presentation of the data: a landing page is provided for each deposited item, but the items are not embedded into a related network of relevant datasets. This might stem from these services' primary focus on long-term preservation. More importantly, in such low-barrier-to-entry data services, metadata schemas associated with data publications are usually limited to a minimum set of information (authors, title, publication date, free text tags, and categories) and inflexible licensing options that can neither fully cover the complex ownership relations in cultural heritage data, nor are sufficient for providing detailed provenance information.

In both cases we face the minimal common denominator problem: minimally flexible and minimally specified metadata schemas serving as a common base for the accommodation of large amount of heterogeneous data will necessarily bring about at least some loss of information that would otherwise enable productive reuse of the dataset. Such limited possibilities for contextualising and documenting data may keep important assumptions, procedures, processes, and decisions that were made at the different stages of data collection and curation hidden from potential re-users of the deposited dataset. As Carlson and Anderson remind us, data are always cooked in specialised ways within each and every research project.[62] Making the steps of this cookery process explicit is especially important when data designed to answer specific research questions are derived from cultural artefacts carrying their own long life-stories and *thick descriptions*.

Recognising these limitations, which are imposed by insufficient metadata and deficient documentation on reuse, highlights an important aspect of successful data management. That is, to make datasets truly reusable, data should achieve autonomy from their curator. In Carlson and Anderson's words: 'Data re-use not only involves the disconnection of data from the people they represent but also from the researchers

61 Hoekstra, Groth, and Charlaganov, 'Linkitup', p. 96.

62 Carlson and Anderson, 'What *Are* Data?', 144; also cited by Poole, 'Now is the Future Now?', para. 20.

who collected them. This opens up the central question as to how data collected or constructed by one researcher can be trusted or even understood by another.'[63]

In the arts and humanities this act of disconnection is a recurring pattern. Artefacts first become separated from their producers (e.g. from the photographer or writer) when they are brought into cultural heritage institutions. The second separation occurs when digital surrogates, descriptions, and other additions to the history, discoverability, and *thick description* of artefacts — in optimal cases at least — step outside the bounds of the cultural heritage institutions responsible for their preservation and digital curation. The third separation occurs when research data derived from these digitally available cultural data is shared and reused, making it available for continuous enrichment and analysis in multiple research contexts. This third separation is a slowly emerging scholarly practice that is facing many economic, technical, institutional, infrastructural, but primarily, and most importantly, cultural barriers. The more support data sharing practices receive, the more important the question is of how to keep these multiple contexts of the *thick descriptions* of cultural data available for continuous analysis and enrichment. Enabling FAIR data management to realise its promises in the arts and humanities requires a mutual understanding between the epistemic cultures of the various stakeholders involved in the co-creation of the scholarly data continuum, ranging from primary sources to multiple reuse cases.

Conclusions: On our Way towards a Truly FAIR Ecosystem for the Arts and Humanities

It is now beyond question that opening up access to scholarly knowledge is a key value of twenty-first-century academia. The paradigm shift towards digital and computational research methods brings about more sustainable, more connected, and community-driven models of scholarly production. Global policies like FAIR data management have a vital role in catalysing and streamlining such innovations, and also in transposing and defining the ways in which research is designed,

63 Carlson and Anderson, 'What *Are* Data?', 643.

performed and evaluated, and the ways in which knowledge is shared. However, in order to embrace the new potentials of computational innovation and to implement high-level principles in a way that will serve the flourishing of the arts and humanities disciplines, there are concerns we need to systematically address first, using focussed activities both from within arts and humanities research, and at the level of open science policies. These include:

1. Data-drivenness is not yet a mature concept in the arts and humanities. Consequently, there is a need for consolidated interpretative frameworks aimed at helping to reach consensus about what can be considered to be research data[64] in the arts and humanities disciplines, and what is not. Furthermore, enhancing data literacy requires the integration of new skills and new professional roles with the arts and humanities higher education curricula.

   On the one hand, the institutional availability of expert data curator staff (librarians, data scientists, and digital humanities experts) who have both subject expertise,[65] and knowledge of digital humanities and data science techniques, is critical for the support of the vernacularisation of FAIR data management skills. On the other hand, we can expect that arts and humanities research institutions will not have equal access to these support services, or will not be ready for their rapid implementation. Therefore, as a more flexible and more inclusive solution, we recommend European research infrastructures complement the efforts of research institutions with widely accessible data management services (such as repository finders)[66]

64 At the same time, we can expect that the en masse application of global FAIR data policies will also have an incremental and large-scale effect on the notion of data in the arts and humanities as researchers will be forced to interpret certain outputs of their research projects as data.

65 Subject expertise and capacity for one-to-one consultancy would be key contributions for aligning disciplinary culture with data management best practices. This could prevent FAIR from being realised merely as a compulsory administrative task of filling in data management templates tailored to the taste of the different funding bodies, or reducing it to a set of technical requirements.

66 The Data Deposit Recommendation Service (DDRS), which has been developed as functional demonstrator within the Humanities at Scale project, an offspring of DARIAH-EU, is a good example of services helping to establish good data

and advocacy activities (webinars, workshops, e-learning materials, collecting, and sharing exemplary case studies). For instance, the translation of science policies (which are often expressed in science-centric language) into widely applicable terms and disciplinary contexts is an important step in preventing humanities researchers from feeling marginalised and disengaged. By uncovering some of the cornerstones for reconciling disciplinary traditions with FAIR data management, this chapter aims to contribute to this translation.

2. In the arts and humanities, data are collected rather than generated. The history of practices determines the way in which cultural resources are made available. Dealing with non-digital heterogeneous materials has many implications for data fluidity and data-reuse.[67] Most importantly, being attentive to knowledge structures submerged beyond the digital horizon is essential, if we are to avoid research being skewed towards easily available, easy to find online resources, generating further enrichment and even greater visibility — but only for this very small fraction of cultural heritage. Such asymmetry and distortion can cause potentially irreparable damage to our understanding of human culture. Building research infrastructures that do not completely isolate data from their source institutions, but rather incorporate traditional archival practices and knowledge, and facilitate mediation and connections between the computational and the analogue epistemic cultures, could help avoid such potential distortions.
3. In the arts and humanities, data show a highly networked but also highly scattered picture. They are networked in the sense that, due to the intertwined systems of knowledge representation and knowledge creation, it is rather difficult to decouple the never-raw source data from the layers of

management practices in arts and humanities. *DDRS*, https://ddrs-dev.dariah.eu/ddrs/

67 Anne Baillot, Michael Mertens, and Laurent Romary, 'Data Fluidity in DARIAH — Pushing the Agenda Forward', *BIBLIOTHEK Forschung Und Praxis*, 39.3 (2015), 350–57, https://doi.org/10.1515/bfp-2016-0039

analysis that have been built upon them. As a result, scholarly data forms a continuum with not always clearly delineable primary-, intermediate-, and result-data components. In current practice, these different layers of analysis are separated by institutional silos, and only in the rarest of cases can they stay connected to each other. Ensuring that this long continuum is kept together from either end poses special challenges in a data management and hosting infrastructure. Establishing a framework that could serve as a general baseline for interactions between scholars, data centres, and heritage institutions will be an essential component of the FAIR data ecosystem in the arts and humanities domain. Such a trusted network of stakeholders could enable all the relevant actors to connect and together improve access to cultural heritage data, making transactions related to the scholarly use of cultural heritage data more visible and transparent.

4. An important feature of computationally mediated research ecosystems is the autonomy of datasets: as shared assets on a technical level, datasets become disconnected from their creators and contexts of creation, yet epistemologically they still remain, to a certain extent, dependent on these creators and contexts of creation. In the arts and humanities, this act of disconnection is a recurring pattern, and ranges from artefacts first becoming separated from their producers through the opening up of cultural heritage (source) data curated by cultural heritage institutions, to sharing research data and making it available for reuse and reanalysis in multiple research contexts. Such multiple separation events have implications not only in terms of the shared ownership of data, but also in terms of knowledge transfer between these different stakeholder groups. As can be seen, there is a critically high risk of losing contextual information around research sources, which is essential for their productive reuse in the course of remediation of scholarly data. The more support data sharing practices receive, the more important the question: how to prevent this loss and how to keep these multiple contexts of the *thick descriptions* of cultural data available for continuous analysis and enrichment?

Enabling FAIR data management to realise its promise in the arts and humanities requires mutual understanding between the epistemic cultures involved in the co-creation of the scholarly data continuum, ranging from the primary sources to multiple reuse cases. Creating a common online environment to support smooth, end-to-end communication between key actors involved in cultural heritage knowledge creation (cultural heritage institutions, data centres, research institutions, individual researchers) where information on the datasets could be published both manually and automatically (e.g., licensing, citation, reuse, enrichments, and contact information for the persons responsible for curation) would be a key step in keeping together the different layers of analysis, and achieving a better alignment of data creation and curation with downstream reuse.

5. Finally, it is rather difficult to have a fair view of findable, accessible, interoperable, and reusable data management in the humanities without considering the actual situation in the domain of publications. Aligning the slowly emerging scholarly practice of data sharing with the inadequately ageing institutions of book and article publishing is a key step towards a more open, more connected, more transparent, and more sustainable research ecosystem.

   Such considerations may pave the way to a better understanding of the discipline-specific challenges in data production and may, therefore, help to realise the promises of the FAIR guidelines in an arts and humanities context. Building a domain-specific data sharing ecosystem will require continuous checks on where the gaps are between the different epistemic cultures, what is hidden, and what remains unknown. Only this can guarantee a truly functioning and sustainable FAIRness, where neither the sunken substructure of the knowledge iceberg, nor the *thick descriptions*, will be lost for good.

Bibliography

Akers, Katherine G., and Jennifer Doty, 'Disciplinary Differences in Faculty Research Data Management Practices and Perspectives', *International Journal of Digital Curation*, 8 (2013), 5–26, https://doi.org/10.2218/ijdc.v8i2.263

Australian FAIR Access Working Group, *Policy Statement on FAIR Access to Australia's Research Outputs*, https://www.fair-access.net.au/fair-statement

Baillot, Anne, 'A Certification Model for Digital Scholarly Editions: Towards Peer Review-Based Data Journals in the Humanities', *HAL* (2016), halshs-01392880, https://halshs.archives-ouvertes.fr/halshs-01392880/document

Baillot, Anne, Laurent Romary, and Michael Mertens, 'Data Fluidity in DARIAH — Pushing the Agenda Forward', *BIBLIOTHEK Forschung Und Praxis*, 39 (2015), 350–57, https://doi.org/10.1515/bfp-2016-0039

Beneš, Jakub, et al., *The CENDARI White Book of Archives* (2016), http://www.cendari.eu/sites/default/files/WhiteBook-Web.pdf

Birnholtz, Jeremy P., and Matthew J. Bietz, 'Data at Work: Supporting Sharing in Science and Engineering', in *Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work*, GROUP '03 ed. Kjeld Schmidt, Mark Pendergast, Marilyn Tremaine and Carla Simone (New York: ACM, 2003), pp. 339–48, https://doi.org/10.1145/958160.958215

Bryant, Mike, et al., 'The EHRI Project — Virtual Collections Revisited', in *Social Informatics*, ed. by Luca Maria Aiello and Daniel McFarland (Cham, Switzerland: Springer International Publishing, 2015), pp. 294–303, https://doi.org/10.1007/978-3-319-15168-7_37

Carlson, Samuelle, and Ben Anderson, 'What *Are* Data? The Many Kinds of Data and their Implications for Data Re-Use', *Journal of Computer-Mediated Communication*, 12 (2007), 635–51, https://doi.org/10.1111/j.1083-6101.2007.00342.x

Cendari, http://www.cendari.eu/

'Publicly Available Research Guides', *Cendari*, http://www.cendari.eu/thematic-research-guides/available-research-guides

Collins, Sandra, et al., *Going Digital: Creating Change in the Humanities* (Berlin: ALLEA E-Humanities Working Group Report, 2015).

DDRS, https://ddrs-dev.dariah.eu/ddrs/

Drucker, Johanna, 'Humanities Approaches to Graphical Display', *Digital Humanities Quarterly*, 5.1 (2011), http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html

Duff, Wendy M., and Verne Harris, 'Stories and Names: Archival Description as Narrating Records and Constructing Meanings', *Archival Science*, 2 (2002), 263–85, https://doi.org/10.1007/BF02435625

Edmond, Jennifer, 'Tradition and Innovation in the Cendari Research Infrastructure', *Review of the National Center for Digitization 25*, ed. by Zoran Ognjanović (Belgrade: Faculty of Mathematics, University of Belgrade, 2015), pp. 2–9.

'Europeana Collections', *Europeana Collections*, https://www.europeana.eu/portal/?locale=en

European Commission, *Digitisation, Online Accessibility and Digital Preservation. Report on the Implementation of Commission Recommendation 2011/711/EU* (2013–2015), http://ec.europa.eu/information_society/newsroom/image/document/2016-43/2013-2015_progress_report_18528.pdf

European Commission, Directorate-General for Research & Innovation, *H2020 Programme Guidelines on FAIR Data Management in Horizon 2020* (26 July 2016), http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

'Guiding Principles for Findable, Accessible, Interoperable and Re-Usable Data Publishing Version B1.0', *FORCE11* (2014), https://www.force11.org/fairprinciples

Geertz, Clifford, *The Interpretation of Cultures*, rev. ed. (New York: Basic Books, 2000).

Harley, Diane, et al., *Assessing the Future Landscape of Scholarly Communication: An Exploration of Faculty Values and Needs in Seven Disciplines* (Berkeley, CA: Center for Studies in Higher Education, 2010), https://escholarship.org/uc/item/15x7385g

Henderson, Margaret E., *Data Management: A Practical Guide for Librarians* (Lanham, MD: Rowman & Littlefield, 2016).

Hodson, Simon, et al., 'Turning FAIR Data into Reality: Interim Report from the European Commission Expert Group on FAIR Data', *Zenodo* (2018), https://doi.org/10.5281/zenodo.1285272

Hoekstra, Rinke, et al., 'An Ecosystem for Linked Humanities Data', in *The Semantic Web*, ed. by Harald Sack, Giuseppe Rizzo, Nadine Steinmetz, Dunja Mladenić, Sören Auer, and Christoph Lange, Lecture Notes in Computer Science (Cham, Switzerland: Springer International Publishing, 2016), pp. 425–40.

Daley, Beth, ed., *Transforming the World with Culture: Next Steps on Increasing the Use of Digital Cultural Heritage in Research, Education, Tourism and the Creative Industries* (The Hague: Europeana Foundation, 2015), https://pro.europeana.eu/files/Europeana_Professional/Publications/Europeana%20Presidencies%20White%20Paper.pdf

Hoekstra, Rinke, Paul Groth, and Marat Charlaganov, 'Linkitup: Semantic Publishing of Research Data', in *Semantic Web Evaluation Challenge*, ed. by Valentina Presutti, Milan Stankovic, Erik Cambria, Iván Cantador, Angelo Di Iorio, Tommaso Di Noia, et al., Communications in Computer and Information Science (Cham Switzerland: Springer International Publishing, 2014), pp. 95–100, https://doi.org/10.1007/978-3-319-12024-9_12

Imming, Melanie, 'FAIR Data Advanced Use Cases: From Principles to Practice in the Netherlands', *Zenodo* (2018), https://doi.org/10.5281/zenodo.1246815

'Iperion Homepage', *Iperion CH*, http://www.iperionch.eu/

Jackson, Virginia, and Lisa Gitelman, 'Introduction', in *'Raw Data' Is an Oxymoron*, ed. by Lisa Gitelman, Geoffrey C. Bowker, and Paul N. Edwards (Cambridge, MA: MIT Press, 2013), pp. 1–14, https://doi.org/10.7551/mitpress/9302.003.0002

Jointly Designing a Data FAIRPORT, Workshop at Lorentz Center@Snellius, Leiden, 13–16 January 2014, https://www.lorentzcenter.nl/lc/web/2014/602/info.php3?wsid=602

Linked Data — Connect Distributed Data across the Web, http://linkeddata.org/

Mohr, Alicia Hofelich, et al., 'When Data is a Dirty Word: A Survey to Understand Data Management Needs Across Diverse Research Disciplines', *Bulletin of the Association for Information Science and Technology*, 42 (2015), 51–53, https://onlinelibrary.wiley.com/doi/full/10.1002/bul2.2015.1720420114

Mostern, Ruth, and Marieka Arksey, 'Don't Just Build It, They Probably Won't Come: Data Sharing and the Social Life of Data in the Historical Quantitative Social Sciences', *International Journal of Humanities and Arts Computing*, 10 (2016), 205–24, https://doi.org/10.3366/ijhac.2016.0170

Moulin, Claudine, et al., *Research Infrastructures in the Digital Humanities* (Strasbourg: European Science Foundation, 2011), http://www.esf.org/fileadmin/user_upload/esf/RI_DigitalHumanities_B42_2011.pdf

Nauta, Gerhard Jan, and Wietske van den Heuvel, *Survey Report on Digitisation in European Cultural Heritage Institutions 2015* (The Hague: DEN Foundation/Europeana/ENUMERATE, 2015), http://enumeratedataplatform.digibis.com/reports/survey-report-on-digitisation-in-european-cultural-heritage-institutions-2015/detail

Poole, Alex H., 'Now is the Future Now? The Urgency of Digital Curation in the Digital Humanities', *Digital Humanities Quarterly*, 7.2 (2013), http://www.digitalhumanities.org/dhq/vol/7/2/000163/000163.html

Priddy, Mike, and Nicola Horsley, 'Deliverable D3.1 Report on Historical Data as Sources', *KPLEX* (2018), https://kplexproject.files.wordpress.com/2018/06/kplex_deliverable-d3-1.pdf

Prost, Hélène, Cécile Malleret, and Joachim Schöpfel, 'Hidden Treasures: Opening Data in PhD Dissertations in Social Sciences and Humanities',

Journal of Librarianship and Scholarly Communication, 3 (2015), https://doi.org/10.7710/2162-3309.1230

Re3data Registry of Research Data Repositories, www.re3data.org

Rice, Robin, and Jeff Haywood, 'Research Data Management Initiatives at University of Edinburgh', *International Journal of Digital Curation*, 6 (2011), 232–44 https://doi.org/10.2218/ijdc.v6i2.199

Rights Statements for in Copyright Objects, http://rightsstatements.org/en/

Sahle, Patrick, and Simone Kronenwett, 'Jenseits der Daten: Überlegungen zu Datenzentren für die Geisteswissenschaften am Beispiel des Kölner "Data Center for the Humanities"', *LIBREAS. Library Ideas*, 23 (2013), https://libreas.eu/ausgabe23/09sahle/

Schöch, Christof, 'Big? Smart? Clean? Messy? Data in the Humanities', *Journal of Digital Humanities*, 2.3 (2013), http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/

Tennant, Jonathan P., 'ResearchGate, Academia.Edu, and Bigger Problems with Scholarly Publishing', *Green Tea and Velociraptors* (2 February 2017), http://fossilsandshit.com/researchgate-academia-edu-and-bigger-problems-with-scholarly-publishing/

—— 'The State of the Art in Peer Review', *FEMS Microbiology Letters*, 365.19 (2018), https://doi.org/10.1093/femsle/fny204

Thoegersen, Jennifer L., '"Yeah, I Guess that's Data": Data Practices and Conceptions among Humanities Faculty', *Libraries and the Academy*, 18 (2018), 491–504.

Warwick, C., et al., 'Evaluating Digital Humanities Resources: The LAIRAH Project Checklist and the Internet Shakespeare Editions Project', in *Openness in Digital Publishing: Awareness, Discovery, and Access. Proceedings of the 11th International Conference on Electronic Publishing,* Vienna, 13–15 June 2007, ed. by Leslie Chan and Bob Martens (Vienna, Austria: ELPUB, 2007), pp. 297–306, https://publik.tuwien.ac.at/files/pub-ar_7877.pdf

Wilkinson, Mark D., et al., 'The FAIR Guiding Principles for Scientific Data Management and Stewardship', *Scientific Data*, 3 (2016), https://doi.org/10.1038/sdata.2016.18

Woeber, Catherine Anne, 'Towards Best Practice in Research Data Management in the Humanities' (unpublished master's dissertation, School of Information Management, Victoria University of Wellington, 2017), http://researcharchive.vuw.ac.nz/handle/10063/6620