The risk of losing thick description: Data management challenges Arts and Humanities face in the evolving FAIR data ecosystem

Erzsébet Tóth-Czifra

The risk of losing thick description: Data management challenges Arts and Humanities face in the evolving FAIR data ecosystem

Erzsébet Tóth-Czifra

Realising the promises of FAIR within discipline-specific scholarly practices

Since their inception in 2014, the FAIR principles[1] have come a long way in serving the global need for generic guidelines for data management and stewardship. Addressing one of the grand challenges of scientific innovation, namely the need for infrastructure supporting the reuse of scholarly data, the FAIR principles have become increasingly influential since their formulation by a wide range of stakeholder groups gathered[2] as a framework for the enhancement and optimisation of the digital ecosystem surrounding scholarly data publication.

The strong need for guidelines enabling and incentivising sustainable, connected, easily accessible and cost-effective models of scholarly data curation was clearly reflected by the FAIR principles' reception. The wide embrace and support of 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 to make their way into global policies worldwide[3]- despite the many open questions their implementation leaves behind and the obvious lack of agreed discipline-level implementation plans and models.

Considering its deep embeddedness into the European scientific innovation and policy landscape, FAIR principles have all the potential to have a huge impact on the future landscape and shape the underlying dynamics of knowledge creation for the better.

[1] Mark D. Wilkinson and others, 'The FAIR Guiding Principles for Scientific Data Management and Stewardship', *Scientific Data*, 2016 <https://doi.org/10.1038/sdata.2016.18>.

[2] 'Lorentz Center - Jointly Designing a Data FAIRPORT from 13 Jan 2014 through 16 Jan 2014' <https://www.lorentzcenter.nl/lc/web/2014/602/info.php3?wsid=602> [accessed 10 September 2018].

[3] See e.g. *H2020 Programme Guidelines on FAIR Data Management in Horizon 2020*, 2016 <http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf> [accessed 10 September 2018].or the 'Policy Statement on FAIR Access to Australia's Research Outputs' 'Australian FAIR Access Working Group' <https://www.fair-access.net.au/fair-statement> [accessed 10 September 2018].

This chance however can easily be missed if the specific dynamics of scientific production in humanities are not addressed in their discipline-level implementation.

With the goal of making FAIR meaningful and helping to realise its promises in an arts and humanities context, in this paper we describe some of the defining aspects underlying the domain-specific epistemic processes that pose hidden or not so hidden challenges in the FAIRification of knowledge creation in arts and humanities. In particular, by applying the FAIR data guiding principles to arts and humanities data curation workflows, it will be uncovered that contrary to their general scope and deliberately domain-independent nature, they have been implicitly designed according to underlying assumptions about how knowledge creation operates and communicates. In the following sections we are addressing three such premises: first, scholarly data or metadata is digital by nature[4], second, scholarly data is always created and therefore owned by researchers[5], and third, there is a wide community-level agreement on what can be considered as scholarly data. The problems around such assumptions in arts and humanities are cornerstones in reconciling disciplinary traditions with FAIR data management. By addressing them one by one, we aim to contribute to the better understanding of discipline-specific needs and challenges in data production, discovery and reuse. These considerations may facilitate the inclusive and optimal implementation of the high-level principles in a way that will serve the flourishing of the arts and humanities disciplines rather than imposing limitations on its epistemic practices.

A cultural knowledge iceberg sunken into an analogue world

A fundamental difference between the epistemic cultures of STEM and arts and humanities is that in the arts and humanities the wide range of scholarly information artefacts, works of art, written documents of all sorts, recordings, annotations etc. broadly referred to as research data (in the sense of Henderson 2016:2)[6] are not

[4] See the Preamble of the principles FORCE11 'Guiding Principles for Findable, Accessible, Interoperable and Re-Usable Data Publishing Version B1.0', *FORCE11*, 2014 <https://www.force11.org/fairprinciples> [accessed 10 September 2018]. where the eScience ecosystem is clearly indicated as the domain of FAIR data management.

[5] Note that in the Preamble there is no reference to data providers and data curators other than researchers (like private or publicly funded providers of medical data or curators of cultural heritage) nor they are mentioned among the stakeholders.

[6] 'Research data is data that is collected, observed, or created, for purposes of analysis to produce original research results.' Margaret E. Henderson, *Data Management: A Practical Guide for Librarians* (Rowman & Littlefield, 2016). Other data definitions in a Humanities context are more restrictive e.g. that of Schöch (2013) (Christof Schöch, 'Big? Smart? Clean? Messy? Data in the Humanities', *Journal of*

autonomous products of research projects but are deeply embedded in the cultural memory of Europe, and 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[7] institutions and ranging from national libraries and archives down to small village museums or administrations, are typically not part of the institutional landscape of academia. In spite of this, the digital research ecosystem poses many challenges connected to the exploration and exploitation of the material or collections they hold, and we do not need to get far in the FAIR acronym to recognise these challenges.

The fact that these cultural sources and their enrichments are not only representations of history but also come themselves with a history 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 cultural resources are made available. According to a Europeana Foundation white paper[8] from 2015, only 10% of the European cultural heritage is digitally available (300 million objects). 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 joint digitisation efforts in Europe[9], these numbers suggest that for the foreseeable future arts and humanities research will retain its hybrid nature encompassing varying degrees of digital and analogue elements, thus calling for both automated and manual workflows and practices.

To take an example for the illustration of how much effort and investment is required to satisfy the basic requirement of data being digital in a cultural heritage context,

Digital Humanities, 2013 <http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/> [accessed 19 July 2018].) As we will note later in this paper, the notion of research data is far from being straightforward in arts and humanities.

[7] Galleries, libraries, archives, museums.

[8] Beth Daley, *Transforming the World with Culture: Next Steps on Increasing the Use of Digital Cultural Heritage in Research, Education, Tourism and the Creative Industries.*, 2015 <https://pro.europeana.eu/files/Europeana_Professional/Publications/Europeana%20Presidencies%20White%20Paper.pdf> [accessed 10 September 2018]. See also the same numbers in fig. 3.6. of the ENUMERATE Survey Report on Digitisation in European Cultural Heritage Institutions Gerhard Jan Nauta and Wietske van den Heuvel, *Survey Report on Digitisation in European Cultural Heritage Institutions* <http://www.den.nl/art/uploads/files/Publicaties/ENUMERATE_Report_Core_Survey_3_2015.pdf> [accessed 10 September 2018].

[9] *Digitisation, Online Accessibility and Digital Preservation. Report on the Implementation of Commission Recommendation 2011/711/EU* <http://ec.europa.eu/information_society/newsroom/image/document/2016-43/2013-2015_progress_report_18528.pdf> [accessed 10 September 2018].

Samuelle Carlson and Ben Anderson[10] refer to two digitisation projects as cases in point. The CurationProject had been aimed at digitising and making available for study records of a collection of more than 750,000 artefacts and 100,000 field photographs collected since 1884, and the AnthroProject where anthropological materials from a range of countries including fieldwork notes, images, maps and texts had been digitised and distributed through an online database and DVDs. In both projects the major challenge was to build a well-structured, searchable database from the rather heterogenous sources and records. This aim could have been realised as a rather long-term goal in both projects: progressive digitisation, curation and systematic documentation took 30 years in the former case and 30 years in the latter.

Taking a step further towards findability, although digitisation is a preliminary first step towards sharing knowledge, it alone doesn't guarantee visibility and accessibility of cultural heritage data outside of the walls of their hosting institutions. The aforementioned Europeana survey reveals[11] that only one third (34%) of the digitised cultural heritage resources are currently available online, with barely 3% of these works is suitable for real creative reuse, meaning, only this 3% has the chance to fulfil the discipline-specific measures of being FAIR.

There are a number of cultural, social, legal, technical and economic reasons explaining this small percentage of truly reusable cultural heritage data. These circumstances impact greatly 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 (but lead to a bad culture of sharing and hinder greatly the technical prerequisites of interoperability)

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 use conditions of cultural heritage sources are embedded. Determining ownership status over research based upon such material poses challenges in many cases, as it is on some level shared between the researcher who carries out scientific analysis on the

[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] Beth Daley. p. 9.

source materials, the institution that hosts and curates this material, and the people and cultures that give rise to the objects in question (e.g. photographers but also subjects on the photographs). Establishing 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 the artefacts) are often embedded into historical practices, in particular in eras or contexts where the legal-ethical framework defining present-day data exchange was either non-existent or irrelevant. Obviously, those handling these data could not know in advance that some information e.g. attribution or consent from the rights holders has to be collected, this requirement was brought about only by the digital age. Tracing back provenance of such records is a time-consuming and difficult process filled with uncertainties and unclarity especially in the case of collections inherited from other institutions.[13]

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

Broad investigations of archival practices conducted within the framework of the Knowledge Complexity (KPLEX) project[14] by Mike Priddy and Nicola Horsley reveal how

[12] To illustrate this complexity, let us cite here two examples Carlson and Anderson (2007:643) quotes from their 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 shows a ritual/secret process. "(...) ''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."

[13] This legal uncertainty in the identification of 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, 4 of of the 12 standardised rights statements refer to unclear legal statuses. These are: In Copyright - Rights-holder(s) Unlocatable or Unidentifiable, Copyright Not Evaluated, Copyright Undetermined, No Known Copyright. 'RightsStatements.Org' <http://rightsstatements.org/en/> [accessed 10 September 2018].

[14] Mike Priddy, Nicola Horsley, *Deliverable D3.1 Report on Historical Data as Sources* <https://kplexproject.files.wordpress.com/2018/06/kplex_deliverable-d3-1.pdf> [accessed 10 September 2018]. KPLEX is a Horizon 2020 project aimed at the investigation of the ways in which a

such legal restrictions affect also technical and cultural aspects of data sharing in the cultural heritage domain. 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 do not impact the barriers to reuse of content only 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 instance, some collection descriptions cannot be made available online because they contain biographical information about the person who donated them.

As the below 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 FAIR recommendation that metadata should be open by default even in the case of sensitive data.[15]

> … these 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 needs to fulfil legal requirements and to avoid penalty risks drive a conservative stance where there may be any uncertain or grey area and incentivises practices of reduced sharing or holding data back out of fear of lawsuits against and legal liability of the respective institutions. The lack of a clear definition of the legal barriers puts a large portion of cultural heritage material onto a minefield neither practitioners in cultural heritage institutions nor scholars are willing to step on. Abandonment of certain research questions or attempts to make sources accessible due to legal uncertainty and the lack of accurate, transparent, and easily understandable conditions

focus on 'big data' in ICT research elides important issues about the information environment we live in. The project focuses of 4 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] 'The basic core is proposed as discovery metadata, persistent identifiers and access to the data or, at minimum, metadata.' Simon Hodson and others, 'Turning FAIR Data into Reality: Interim Report from the European Commission Expert Group on FAIR Data', 2018, p. 57 <https://doi.org/10.5281/zenodo.1285272>.

[16] Mike Priddy, Nicola Horsley, p. 65.

of access to the documents is an even bigger obstacle to FAIRification in the cultural heritage domain than the institution of legal protection which it aims to serve.

Case study: Removal of photos from the archival research guides of the CENDARI project due to the lack of information on their reuse conditions

The following case study from the project CENDARI[17] illustrates how legal, cultural, and data-management dimensions of intransparency can lock away valuable and relevant cultural data from being 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 the 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 from the publication.

The online catalogues for the sources neither gave rights holder information or contact for publication permission nor indicated 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 the lack of transparency and proper data management that, originating from external data providers, is out of control of the researcher community. If the longevity of cultural heritage data is defined by their presence in scientific, cultural and social discourses, once we lose access to its reuse conditions, we lose them entirely.

The risk of losing the *thick description* upon the remediation of cultural heritage

[17] 'Cendari' <http://www.cendari.eu/> [accessed 10 September 2018].

[18] 'Available Research Guides | Cendari' <http://www.cendari.eu/thematic-research-guides/available-research-guides> [accessed 10 September 2018].

The advent of digital research infrastructures opened up a radically new frontier for the interactions with cultural heritage 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 have been created with different scopes and sizes, like Europeana[19], IPERION CH[20] or CENDARI[21]. They all are aimed at the mission of building bridges, interlinks and networks (e.g. co-referencing systems, conceptual models, ontologies, semantic web frameworks) across different types of resources and institutions, 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 for having played a catalysator role in the development, promotion and implementation of shared protocols and standards (like the Linked Open Data paradigm[22] in arts and humanities) to guarantee the interoperability between heterogeneous data resources. Papers reporting data collection procedures for the research infrastructure projects EHRI[23] and CENDARI[24] give 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, standardisation of shared metadata has brought not only technical and financial challenges. The new ways in which cultural resources have been made available as a part of global networks haven't left the systems of discovery and

[19] 'Europeana Collections', *Europeana Collections* <https://www.europeana.eu/portal/?locale=en> [accessed 10 September 2018].

[20] 'Home', *Iperion CH* <http://www.iperionch.eu/> [accessed 10 September 2018].

[21] 'Cendari'.

[22] 'Linked Data | Linked Data - Connect Distributed Data across the Web' <http://linkeddata.org/> [accessed 10 September 2018].

[23] Mike Bryant and others, 'The EHRI Project - Virtual Collections Revisited', in *Social Informatics*, ed. by Luca Maria Aiello and Daniel McFarland (Springer International Publishing, 2015), pp. 294–303.

[24] Jakub Beneš, Nataša Bulatović, Jennifer Edmond, Milica Knežević, Jörg Lehmann, Francesca Morselli, Andrei Zamoisk, 'The CENDARI White Book of Archives', 2016 <http://www.cendari.eu/sites/default/files/WhiteBook-Web.pdf> [accessed 10 September 2018].

knowledge creation unaffected either. 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 of the computational turn.

One of them has to do with losing control over remediated records of archival knowledge and its complexity. In the course of traditional interactions like in-person visits or one-on-one consultations, archivists had the possibility to freely guide the researcher through the collections and transfer all knowledge relevant 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 interpretation of specific datasets that had been harvested from institutions. Practitioners' concerns about misinterpretations and misuse of the data they carefully curated had been clearly and repeatedly indicated in the interviews.[26]

A speciality[27] of data management in arts and humanities therefore is that it is highly dependent on external data providers, that is, the cultural heritage institutions. As it was touched upon in the CENDARI case study above as well, due to this dependence certain aspects of data management and FAIRification efforts remain out of control of the 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, its curators and its wider provenance), collection descriptions that are part of the standardised and aggregated metadata remain the only reference points to the long history of records.

Creating descriptions is, therefore, a pivotal process but also a complex task. Practitioners showed awareness of how much preparing these online representations and aligning the richest possible descriptions with their limited surface is an interpretative practice. As it has also been pointed out by Wendy Duff and Verne Harris[28], personal decisions made in the course of this knowledge transfer are inherently biased and therefore will foreground certain pieces of information[29] while

[25] www.kplex-project.eu

[26] Mike Priddy, Nicola Horsley, pp. 52–53, 64–68.

[27] However, arts and humanities are not the only disciplines being dependent from external data providers, see e.g. medial and health care studies.

[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,

leaving others sunken in the analogue practices and tacit knowledge. One thing is certain, however: the separation of data from the curators bearing this knowledge and providing instead a thusly impoverished form of online access to such remediated knowledge representations necessarily leads to limitations in conveying their complexity and simulacra that are misleading in their apparent completeness. And this is crucial because the loss of information is the loss of 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 hard work of curators, just as it may impoverish researchers' interpretation and understanding of possible uses of sources. In other words, hiddenness and the loss of the thick description[30] of holdings is part of the story of making the historical and cultural records available for digital and computational discovery. Researchers in arts and humanities always need multiple sources to verify interpretations, but that requires deep knowledge or 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 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 and lost in a computationally mediated context of discovery. Practitioners' concerns about the undigitised or offline base of the knowledge iceberg being forgotten and 'buried at deeper levels of accessibility during this transitional period'[31] were clearly articulated in the KPLEX interviews. It is a serious threat that a new generation of scholars might lose this awareness of materials and knowledge structures sunken behind the digital horizon (that is, one has to know what one cannot find). The main jeopardy of this

while the complexity that distinguishes the unique value of items may be hidden.' Mike Priddy, Nicola Horsley, 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 20th-century 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*, 2000th Revised ed. (New York: Basic Books, 1977), 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] Mike Priddy, Nicola Horsley, p. 79.

effect is that it may skew research towards what's easily available, easy to find and ideally available freely online and generate 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,[32] such distortion effects are also arising from the fact that contrary to the essentially transnational nature of historical research, the digitisation of cultural heritage was largely founded and continue to be funded along national lines and not every country or institution has access to the same resources. This results in substantial differences in the digital and online footprint of different institutional holdings: wealthier institutions might have 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]

We believe that amid the 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 arts and humanities. Keeping an open eye and critical reflection on overall progress as well as limited or immature cases of openness might help in identifying phenomena and situations where the principles enshrined in the first 2 FAIR letters, findability and accessibility, come into conflict with the last one, reusability. If we want to play it right in a 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 behind the digital horizon would be an unreasonable price to pay for the sake of paradigm shift. Awareness of them is a guarantee that we will not have to pay this price and can realise the promises of innovative revolution to the full to enable new forms of scholarly insight and communication.

The scholarly data continuum

[32] Jennifer Edmond, 'Tradition and Innovation in the Cendari Research Infrastructure', *Review of the National Center for Digitization*, 2015, 2–9.

[33] Edmond, p. 4.

The previous sections showcased that in contrast to hard sciences, initial data in arts and humanities is *collected*[34] rather than *generated*[35], 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 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 – ranging from decisions on what to preserve and what not, the choice of classification systems and metadata schemas, the way in which texts and artefacts are photographed or in which text corpora are transcribed and encoded or OCR is corrected –, all of which impose a perspective on and thus influence 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's 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 gives access to and enables other layers of analysis and knowledge creation resulted by the scholarly activities.

In the current practice these different layers of analysis are separated by institutional silos and only in the rarest cases can they stay connected with each other. As a result,

[34] This distinction and its epistemological consequences are also articulated in Johanna Drucker's study on *capta* versus *data* where capta is "taken" (cf. the term *capta* stems from the Latin word for 'to take'), constructed and is rooted in co-dependent relation between the observer and the experience, while data represent observer-independent models of knowledge given as a natural representation of pre-existing fact. Johanna Drucker, 'Humanities Approaches to Graphical Display', *Digital Humanities Quarterly*, 005.1 (2011).

[35] Claudine Moulin and others, *Research Infrastructures in the Digital Humanities* (Strasbourg: European Science Foundation, 2011), p. 5 <http://darhiv.ffzg.unizg.hr/id/eprint/1888/> [accessed 19 July 2018].

[36] See discussion on the 'fuzzy, implicitly highly networked data' in Humanities that questions the separability of the primary data - intermediate data - result data areas also 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*, 2013 <https://libreas.eu/ausgabe23/09sahle/> [accessed 19 July 2018]. Patrick Sahle and Simone Kronenwett argue that by digitizing 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 on the first place is a general and major challenge in sustainable in reproducible data management and is a topic that deserves more detailed discussion that is could receive within the framework of the present paper.

[37] Virginia Jackson and others, *'Raw Data' Is an Oxymoron*, ed. by Lisa Gitelman, Geoffrey C. Bowker, and Paul N. Edwards (Cambridge, Mass.: The MIT Press, 2013).

the actual continuum in the knowledge creation procedures of the cultural heritage domain is barely reflected in its infrastructure and data management practices.

A key recommendation of the FAIR principles aiming to facilitate access to research data is that data should be stored in trusted and sustainable digital repositories.[38] Taking a 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[39] 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 (1132 result), but also compared to the sibling disciplinary group social and behavioural sciences (331 results). The low number of repositories suggests lower demand for data sharing services or at least a less established data sharing culture in arts and humanities than in other fields of study.[40] On the other hand, however, several recent studies[41] herald the increasing interest in data sharing in arts and humanities at a global disciplinary scale. For instance, in Ruth Mostern and Marieka Arkskey's 2016 study[42] surveying the target users of the Collaborative for Historical Information and Analysis (CHIA) database, 94 per cent of the respondents indicated that they would consider putting their data in a repository.[43]

[38] Hodson and others, 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 and others, Communications in Computer and Information Science (Springer International Publishing, 2014), pp. 95–100; Sandra Collins and others, 'Going Digital: Creating Change in the Humanities' (unpublished PhD Thesis, ALLEA, 2015).

[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>.

[43] This seems significant progress e.g. over Diane Harley et al.'s study where evidence is shown that historians are cautious about sharing work publicly until it is well-polished. Similarly to many other fields in 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 criticized 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." Diane

Understanding this wide gap between intentions vs. real willingness vs. 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?

Of course, sharing data necessarily implies having/owning data. In addition to the aforementioned complexities in shared ownership of primary sources that forms a major hindrance to data sharing, having data or working with data is not always a straightforward concept, especially in the traditional fields of arts and humanities. Iterated and large-scale surveys would be beneficial to assess whether and to what extent the term data is still a dirty word[44] 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.

Surveys from the past 5 years[45] reveal a great deal of uncertainty in arts and humanities researchers' conceptions of data and its applicability to their own work.[46] Concerns and difficulties around the concept of data were clearly reflected in responses of 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 University of Nebraska-Lincoln were

Harley and others, 'Assessing the Future Landscape of Scholarly Communication: An Exploration of Faculty Values and Needs in Seven Disciplines', 2010, p. 451 <https://escholarship.org/uc/item/15x7385g> [accessed 10 September 2018].

[44] Reference to Alicia Hofelich Mohr et al.'s 2015 article Alicia Hofelich Mohr and others, '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://doi.org/10.1002/bul2.2015.1720420114>.

[45] Akers and Doty; Mohr and others; Hélène Prost, Cécile Malleret, and Joachim Schöpfel, 'Hidden Treasures: Opening Data in PhD Dissertations in Social Sciences and Humanities', *Journal of Librarianship and Scholarly Communication*, 3.2 (2015) <https://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, p. 492. 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.

[47] Thoegersen Jennifer L.

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 doesn’t specify 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 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, information throughputs and outputs (laser scanner data, musical notations, voice recordings, annotations, critical editions etc.) produced by the wide range disciplines that are both standing under the umbrella term 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 the shift in paradigm towards data-driven and computational methods across the whole disciplinary range of arts and humanities. Yet, there are still plenty of interrelated issues that prevent data sharing in subject repositories (which are, as we have seen, central data services in the implementation of FAIR principles) and hamper reuse from 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 Arskey[49] capture many of such interrelated problems that define the current repository landscape in arts and humanities, lingering 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 arts and humanities, keeping in mind that they are not equally plagued with the described problems.

As it had been pointed out in several other discipline-specific data management studies[50] there is a lack of incentives and rewards to dedicate considerable amount of

[48] Jennifer L. Thoegersen, p. 501.

[49] Mostern and Arksey.

[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*, 007.2 (2013); Catherine Anne Woeber, ‘Towards Best Practice in Research Data Management in the Humanities’ (School of Information Management, Victoria University of Wellington, 2017) <http://researcharchive.vuw.ac.nz/handle/10063/6620> [accessed 10 September 2018].

time, effort, and expertise to prepare data for computational analysis and make it compliant with the standards and data model of the repositories. 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 them or contribute to them.[51] In addition, repository developers and standardisation bodies then do not receive a significant enough base of input 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 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 slightly visible part of the wider disciplinary communities, preventing their work and approach from being accessible and strongly represented to students and peers to encourage them to share their data too and thus at the end the strongest appeal of the use of repositories is not able to work its charm.

Having been inspired by the 2003 study of Jeremy P. Birnholtz and Mattew J. Bietz[52], Mostern and Arskey describe this complex phenomenon as the lack of social life of data. Recognising the importance of a community aspect around robust data sharing culture, that is, documents and deposited datasets are not only means to deliver information, but they are also meant to maintain social groups and exchange around them, they came to the important conclusion that repositories can only succeed as

[51] Note that guaranteeing the presence of target audience via reaching a critical mass of content was the recipe for success of the two, even nowadays commonly used academic sharing and networking platforms, ResearchGate and Academia. We can learn a lot from the failures that are underlying their conceptual design and what had become visible only after they reached a critical level of user engagement. Although the original aim of both platforms were helping researchers going beyond paywalls and increasing 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 copy rights which article versions are allowed to be legally shared, low awareness of the importance of licensing issues, support to freemium business models based on selling data on user behaviours) on such a massive scale that 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*, 2017 <http://fossilsandshit.com/researchgate-academia-edu-and-bigger-problems-with-scholarly-publishing/> [accessed 10 September 2018].

[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 (New York, NY, USA: ACM, 2003), pp. 339–348 <https://doi.org/10.1145/958160.958215>.

long as scholarly communities create social communities around them.[53] 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 to build user confidence, as 70% 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 closer the scholarly practices of data depositing and data peer review is also expressed in a checklist of recommendations of the LARIAH project. According these recommendations, the ideal digital resource 1. should have access to good technical support, ideally from a centre of excellence in digital humanities 2. should recruit staff who have both subject expertise and knowledge of digital humanities techniques 3. should also retain this expert staff via having constant access to funds.[54]

Data peer review along these lines, that is, focusing on support and joint development of transparent and good quality data creation without the power dynamics and gatekeeping function that are causing serious challenges in the institution of traditional article and book peer review[55], could also be interpreted as a significant contribution to a more sustainable and more inclusive culture of research evaluation in general. At the same time however, point 3 of the LARIAH recommendations also indicates the serious sustainability challenges of such models in terms of funding. The ability to maintain a both technically and disciplinary highly skilled expert staff around repositories who have the capacity of providing thorough evaluation of a massive number of data deposits that can be expected as a result of FAIR policies doesn't seem

[53] These observations show congruency with the main findings of a much earlier study on the uptake and use of digital resources in arts and humanities, namely LARIAH project (Log analysis of Internet Resources in the Arts and Humanities; see a project description in C. Warwick and others, '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, June 13-15, 2007.*, ed. by Leslie Chan and Bob Martens (Vienna, Austria: ELPUB, 2007), pp. 297–306 <http://elpub.architexturez.net/doc/oai-elpub.id-144_elpub2007> [accessed 10 September 2018].) The project was based at UCL's School of Library Archive and Information Studies and was aimed at identifying different 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 the 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. provided a checklist of recommendations to facilitate both the successful design of digital infrastructures and the recognition and culture around them.

[54] Warwick and others, pp. 302–3.

[55] See e.g. Jonathan P. Tennant, 'The State of the Art in Peer Review', *FEMS Microbiology Letters* <https://doi.org/10.1093/femsle/fny204>.

to be a viable option. As a potential alternative, institutional data stewards[56] and data centers like the Leiden University Centre for Digital Humanities (LUCDH)[57] could at least partially fulfil this role.

An additional challenge in facilitation the culture of data evaluation in arts and humanities is that as it was pointed out by others[58], the scholarly practice of data peer review is still lagging way behind the traditional paradigm of research article publishing serving as the highest value currency of academia. 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

To avoid having deposited datasets being buried into isolated 'data tombs' and to increase the social life of data via making it interoperable and connectible with other data sources, relying on domain-relevant community standards is critical. Achieving compliance with metadata standards is a prerequisite of 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 Humanities.

Aligning the application and the use of repository standards with the long history of data curation can not always be achieved without making compromises, however. In some cases, enforcing commitment to shared standards can lead to a similar loss of detail and information as could be seen in the context of the aggregation of

[56] Rec. 13 of the FAIR Data Action Plan (Hodson and others, p. 73.) recommends to develop two cohorts of professionals to support FAIR data: data scientists embedded in those research projects which 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). The case study published in a recent collection of FAIR data advanced use cases from the Netherlands gives an insight on 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, 23 April 2018), pp. 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', 2018 <https://halshs.archives-ouvertes.fr/halshs-01392880/document> [accessed 10 September 2018].

standardised and machine-interoperable metadata from cultural heritage institutions. In their 2014 and 2016 studies[59] Rinke Hoekstra and his co-authors investigate data sharing practices in humanities and their compliance with linked discovery context. They identify two cases when the risk of losing provenance information is especially high.

First, when data is deposited in discipline-specific but bigger 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[60] 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 making it difficult for them to make share their research in a sustainable way.

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 like Easy, Dryad, Dataverse and Figshare have been created. 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 the citability of data via persistent identifiers, and also guarantee long-term archival storage. On the other hand, however, as argued in the earlier study[61], these generic-scope data sharing platforms bear hidden limitations for discoverability and productive reuse. The first limitation is a result of the rather isolated presentation of data, that is, a landing page is provided for each deposited item but they 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 neither can 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, that is, minimally flexible and specified metadata schemas serving as a common base for the accommodation of large number of heterogeneous data necessarily brings about at least some loss of information that would otherwise enable productive reuse of the

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

[60] Hoekstra and others, p. 426.

[61] Hoekstra, Groth, and Charlaganov, p. 96.

dataset. Such limited possibilities for contextualising and documenting data may keep important assumptions, procedures, processes, and decisions made in the different stages of data collection and curation hidden from potential reusers of the deposited dataset. As Carlson and Anderson[62] remind us, data are always cooked in specialised ways within each and every research project. 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 description*s.

Recognising these limitations, 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 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 arts and humanities, this act of disconnection is a recurring pattern. Artefacts first became separated from their producers (e.g. from their photographer or writer) when making their way into cultural heritage institutions. In a second round, digital surrogates, descriptions and other additions to the history, discoverability and thick description of artefacts are, in optimal cases at least, stepping outside of the walls of the cultural heritage institutions responsible for their preservation and digital curation. The third separation, sharing and reusing research data derived from these digitally available cultural data and making it available for continuous enrichment and analysis in multiple research contexts is yet a slowly emerging scholarly practice, 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 description*s of cultural data available for continuous analysis and enrichment. Enabling FAIR data management to realise its promises in arts and humanities requires mutual understanding between the epistemic cultures of various stakeholders involved in the co-creation of the scholarly data continuum ranging from the primary sources to multiple reuse cases.

Conclusions: On our way towards a truly FAIR ecosystem for the arts and humanities

[62] Carlson and Anderson, p. 144., also cited by Poole, para. 20.

[63] Carlson and Anderson, p. 643.

It is now beyond question that opening up access to scholarly knowledge is a key value of the academy of the 21st century. The paradigm shift towards digital and computational research methods bring about more sustainable, more connected and community-driven models of scholarly production. Global policies like FAIR data management has a vital role in catalysing and streamlining such innovations, to transpose but also define the ways in which research is designed, performed and evaluated, and knowledge is shared. But 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 a lot of questions we need to systematically address first with focussed activities both from within arts and humanities and at the level of Open Science policies. These include:

1. Data-drivenness is not yet a mature concept in arts and humanities. Consequently, there is a need for consolidated interpretative frameworks aimed at helping to reach consensus about what can be considered as research data[64] and what is not in the arts and humanities disciplines and what new skills, professional roles do we need to support and sustain to make data meaningful in scholarship.

Concerning support in vernacularising FAIR data management skills, on the one hand, the institutional availability of expert data curator staff (librarians, data scientists or digital humanities experts) who have both subject expertise[65] and knowledge of digital humanities and data science techniques is critical. On the other hand, however, we can expect that arts and humanities research institutions will not have equal access to these support services or will not be enabled for their rapid implementation. Therefore, as a more flexible and more inclusive solution, we recommend European research infrastructures to complement the efforts of research institutions with widely accessible data management services (like repository finders[66]) and advocacy activities (webinars, workshops, e-learning materials, collecting and sharing exemplary case studies). For instance, the translation of science policies (that are often expressed in

[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 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 to be realized 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), that has been developed as functional demonstrator within the Humanities at Scale project, an offspring of DARIAH-EU, is a good example for services helping to establish good data management practices in arts and humanities. 'DDRS' <https://ddrs-dev.dariah.eu/ddrs/> [accessed 10 September 2018].

science-centric language) into widely applicable terms and disciplinary contexts is an important step in preventing humanities researchers to feel marginalised and disengaged. By uncovering some of the cornerstones in reconciling disciplinary traditions with FAIR data management, the present paper aimed to contribute to this translation.

2. Data in arts and humanities are rather collected than generated. The history of practices determines the way cultural resources are made available. Dealing with non-digital heterogeneous materials has many implications for data fluidity[67] and data-reuse. Most importantly, maintaining a watchful awareness towards knowledge structures sunken behind the digital horizon is essential if we want to avoid research being skewed towards easily available, easy to find online resources, generating further enrichment and even greater visibility, but only of 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 don't 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 avoiding such potential distortions.

3. Data in arts and humanities 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 analysis having been built upon them. As a result, scholarly data forms a continuum with not always clearly delineable primary data - intermediate data - result data components. In the current practice these different layers of analysis are separated by institutional silos and in the rarest cases can they stay connected with each other. Keeping together this long continuum 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 improve together access to cultural heritage data and make transactions related to the scholarly use of cultural heritage data more visible and transparent.

[67] Laurent Romary, Michael Mertens, and Anne Baillot, 'Data Fluidity in DARIAH – Pushing the Agenda Forward', *BIBLIOTHEK Forschung Und Praxis*, 39.3 (2015), 350–357 <https://doi.org/10.1515/bfp-2016-0039>.

4. An important feature of computationally mediated research ecosystem is the autonomy of datasets, that is, as shared assets on the technical level, they become disconnected from their creators and contexts of creation, still, epistemologically they remain to a certain extent dependent on them. In arts and humanities, this act of disconnection is a recurring pattern ranging from artefacts first becoming separated from their producers through opening up 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 not only implications in terms of shared ownership of data but also in terms of knowledge transfer between these different stakeholder groups. As can be seen, the risk of losing contextual information around research sources that are essential for their productive reuse in the course of remediation of scholarly data is more than high. The more support data sharing practices receive, the more important the question is 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 arts and humanities requires mutual understanding between these 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 to 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 on 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 discipline-specific challenges in data production and therefore may help to realise the promises of FAIR guidelines in an arts and humanities context. Building a domain-specific data sharing ecosystem will require continuously checking on where are the gaps are between the different epistemic cultures, what is hidden, what remains unknown. Only this can guarantee a truly functioning and sustainable FAIRness where neither the sunken base of the knowledge iceberg nor thick descriptions will be lost for good.

Acknowledgements

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

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' <https://www.fair-access.net.au/fair-statement> [accessed 10 September 2018]

'Available Research Guides | Cendari' <http://www.cendari.eu/thematic-research-guides/available-research-guides> [accessed 10 September 2018]

Baillot, Anne, 'A Certification Model for Digital Scholarly Editions: Towards Peer Review-Based Data Journals in the Humanities', 2018 <https://halshs.archives-ouvertes.fr/halshs-01392880/document> [accessed 10 September 2018]

Beth Daley, *Transforming the World with Culture: Next Steps on Increasing the Use of Digital Cultural Heritage in Research, Education, Tourism and the Creative Industries.*, 2015 <https://pro.europeana.eu/files/Europeana_Professional/Publications/Europeana%20Presidencies%20White%20Paper.pdf> [accessed 10 September 2018]

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 (New York, NY, USA: ACM, 2003), pp. 339–348 <https://doi.org/10.1145/958160.958215>

Bryant, Mike, Linda Reijnhoudt, Reto Speck, Thibault Clerice, and Tobias Blanke, 'The EHRI Project - Virtual Collections Revisited', in *Social Informatics*, ed. by Luca Maria Aiello and Daniel McFarland (Springer International Publishing, 2015), pp. 294–303

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/> [accessed 10 September 2018]

Collins, Sandra, Natalie Harrower, Dag Trygve Truslew Haug, Beat Immenhauser, Gerhard Lauer, Tito Orlandi, and others, 'Going Digital: Creating Change in the Humanities' (unpublished PhD Thesis, ALLEA, 2015)

'DDRS' <https://ddrs-dev.dariah.eu/ddrs/> [accessed 10 September 2018]

Digitisation, Online Accessibility and Digital Preservation. Report on the Implementation of Commission Recommendation 2011/711/EU, 2015 <http://ec.europa.eu/information_society/newsroom/image/document/2016-43/2013-2015_progress_report_18528.pdf> [accessed 10 September 2018]

Drucker, Johanna, 'Humanities Approaches to Graphical Display', *Digital Humanities Quarterly*, 005 (2011)

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*, 2015, 2–9

'Europeana Collections', *Europeana Collections* <https://www.europeana.eu/portal/?locale=en> [accessed 10 September 2018]

FORCE11, 'Guiding Principles for Findable, Accessible, Interoperable and Re-Usable Data Publishing Version B1.0', *FORCE11*, 2014 <https://www.force11.org/fairprinciples> [accessed 10 September 2018]

Geertz, Clifford, *The Interpretation Of Cultures*, 2000th Revised ed. (New York: Basic Books, 1977)

Gerhard Jan Nauta and Wietske van den Heuvel, *Survey Report on Digitisation in European Cultural Heritage Institutions 2015*, 2015 <http://www.den.nl/art/uploads/files/Publicaties/ENUMERATE_Report_Core_Survey_3_2015.pdf> [accessed 10 September 2018]

H2020 Programme Guidelines on FAIR Data Management in Horizon 2020, 2016 <http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf> [accessed 10 September 2018]

Harley, Diane, Sophia Krzys Acord, Sarah Earl-Novell, Shannon Lawrence, and C. Judson King, 'Assessing the Future Landscape of Scholarly Communication: An Exploration of Faculty Values and Needs in Seven Disciplines', 2010 <https://escholarship.org/uc/item/15x7385g> [accessed 10 September 2018]

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

Hodson, Simon, Sarah Jones, Sandra Collins, Françoise Genova, Natalie Harrower, Leif Laaksonen, and others, 'Turning FAIR Data into Reality: Interim Report from the European Commission Expert Group on FAIR Data', 2018 <https://doi.org/10.5281/zenodo.1285272>

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, and others, Communications in Computer and Information Science (Springer International Publishing, 2014), pp. 95–100

Hoekstra, Rinke, Albert Meroño-Peñuela, Kathrin Dentler, Auke Rijpma, Richard Zijdeman, and Ivo Zandhuis, '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 (Springer International Publishing, 2016), pp. 425–40

'Home', *Iperion CH* <http://www.iperionch.eu/> [accessed 10 September 2018]

Jackson, Virginia, Daniel Rosenberg, Travis D. Williams, Kevin R. Brine, Mary Poovey, Matthew Stanley, and others, *'Raw Data' Is an Oxymoron*, ed. by Lisa Gitelman, Geoffrey C. Bowker, and Paul N. Edwards (Cambridge, Mass.: The MIT Press, 2013)

Jakub Beneš, Nataša Bulatović, Jennifer Edmond, Milica Knežević, Jörg Lehmann, Francesca Morselli, Andrei Zamoisk, 'The CENDARI White Book of Archives', 2016 <http://www.cendari.eu/sites/default/files/WhiteBook-Web.pdf> [accessed 10 September 2018]

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

Jonathan P. Tennant, 'ResearchGate, Academia.Edu, and Bigger Problems with Scholarly Publishing..', *Green Tea and Velociraptors*, 2017 <http://fossilsandshit.com/researchgate-academia-edu-and-bigger-problems-with-scholarly-publishing/> [accessed 10 September 2018]

'Linked Data | Linked Data - Connect Distributed Data across the Web' <http://linkeddata.org/> [accessed 10 September 2018]

'Lorentz Center - Jointly Designing a Data FAIRPORT from 13 Jan 2014 through 16 Jan 2014' <https://www.lorentzcenter.nl/lc/web/2014/602/info.php3?wsid=602> [accessed 10 September 2018]

Melanie Imming, *FAIR Data Advanced Use Cases: From Principles to Practice in the Netherlands* (Zenodo, 23 April 2018) <https://doi.org/10.5281/zenodo.1246815>

Mike Priddy, Nicola Horsley, *Deliverable D3.1 Report on Historical Data as Sources* <https://kplexproject.files.wordpress.com/2018/06/kplex_deliverable-d3-1.pdf> [accessed 10 September 2018]

Mohr, Alicia Hofelich, Josh Bishoff, Carolyn Bishoff, Steven Braun, Christine Storino, and Lisa R. Johnston, '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://doi.org/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, Julianne Nyhan, Arianna Ciula, Margaret Kelleher, Elmar Mittler, Marko Tadić, and others, *Research Infrastructures in the Digital Humanities* (Strasbourg: European Science Foundation, 2011) <http://darhiv.ffzg.unizg.hr/id/eprint/1888/> [accessed 19 July 2018]

Poole, Alex H., 'Now Is the Future Now? The Urgency of Digital Curation in the Digital Humanities', *Digital Humanities Quarterly*, 007 (2013) <http://www.digitalhumanities.org/dhq/vol/7/2/000163/000163.html>

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>

'RightsStatements.Org' <http://rightsstatements.org/en/> [accessed 10 September 2018]

Romary, Laurent, Michael Mertens, and Anne Baillot, 'Data Fluidity in DARIAH – Pushing the Agenda Forward', *BIBLIOTHEK Forschung Und Praxis*, 39 (2015), 350–357 <https://doi.org/10.1515/bfp-2016-0039>

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*, 2013 <https://libreas.eu/ausgabe23/09sahle/> [accessed 19 July 2018]

Schöch, Christof, 'Big? Smart? Clean? Messy? Data in the Humanities', *Journal of Digital Humanities*, 2013 <http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/> [accessed 19 July 2018]

Tennant, Jonathan P., 'The State of the Art in Peer Review', *FEMS Microbiology Letters* <https://doi.org/10.1093/femsle/fny204>

Warwick, C., M. Terras, I. Galina, P. Huntington, and N. Pappa, '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, June 13-15, 2007.*, ed. by Leslie Chan and Bob Martens (Vienna, Austria: ELPUB, 2007), pp. 297–306 <http://elpub.architexturez.net/doc/oai-elpub.id-144_elpub2007> [accessed 10 September 2018]

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, and others, 'The FAIR Guiding Principles for Scientific Data Management and Stewardship', *Scientific Data*, 2016 <https://doi.org/10.1038/sdata.2016.18>

Woeber, Catherine Anne, 'Towards Best Practice in Research Data Management in the Humanities' (School of Information Management, Victoria University of Wellington, 2017) <http://researcharchive.vuw.ac.nz/handle/10063/6620> [accessed 10 September 2018]