- Preprint of submission to the CODATA Data Science Journal, version as of May 2019 –

Reconciling the Cultural Complexity of Research Data: Can we Make Data Interdisciplinary without Hiding Disciplinary Knowledge?

Michelle Doran, Jennifer Edmond, Georgina Nugent-Folan

Centre for Digital Humanities, Trinity College Dublin, Dublin, Ireland.

Corresponding author's email: doranm1@tcd.ie

Author Information

Michelle Doran is Research Assistant and Project Officer at the Centre for Digital Humanities, Trinity College Dublin and is presently contributing to the Horizon 2020 Knowledge Complexity (KPLEX) project (www.kplex-project.eu). She a member of the Advisory Board of HubIT, the HUB for boosting the Responsibility and Inclusiveness of ICT enabled Research and Innovation through constructive interactions with Social Sciences and Humanities research.

Jennifer Edmond is Director of Strategic Projects for the Faculty of Arts, Humanities and Social Sciences at Trinity College Dublin. She has a significant profile in European research and research policy circles, having coordinated or partnered in a number of significant research and infrastructure projects. Jennifer is a member of the Board of Directors of the DARIAH ERIC, a body she represents on the European Commission's Open Science Policy Platform (OSPP).

Georgina Nugent-Folan is Postdoctoral Research Fellow at the Centre for Digital Humanities, Trinity College Dublin and is presently contributing to the Horizon 2020 Knowledge Complexity (KPLEX) project (www.kplex-project.eu).

Abstract

One of the major factors inhibiting interdisciplinary data driven research is how to capture provenance and facilitate the discovery, use, and reuse of discipline specific research 'data'. The growing pressure to find ways to enable technological and cultural compliance with the nascent European Open Science Cloud (EOSC) is intensifying what is an already difficult proposition. This paper outlines the preliminary findings of the Horizon2020-funded Knowledge Complexity (KPLEX) project (https://kplex-project.eu/), which is investigating the delimiting effect digital mediation and datafication can have on rich, complex cultural data. As an ICT-35 'sister project', KPLEX is humanities-led and ICT-funded and the project partners are approaching this challenge in a comparative, multidisciplinary and multisectoral fashion. This paper investigates and outlines the wider lessons to be learned from data reuse practices in the humanities, fields well accustomed to dealing with complex, hybrid, and noisy data. Understanding the challenges that complex research data pose to our research infrastructures will help facilitate the transition to a truly open and interdisciplinary frontier for scientific research and innovation. This paper focuses on the criteria and guidelines for the recognition and classification of data, both between and within research data infrastructures and research communities. It describes the ways in which the effectiveness of traditional signposting mechanisms such as research classification systems, taxonomies, and metadata are weakened in the digital age and sets forth the implications that these forces will have for a wider set of disciplines under the vision of the EOSC.

Keywords

Data Reuse, Interdisciplinarity, Metadata, Digital Curation, European Open Science Cloud, Digital Humanities.

> Even if one were to accept the fiction of the universal database managed by a single authority, the fundamental problem of meaningfully, and predictably, parsing that archive remains (Raley 2013:129).

1. Introduction.

On 26 October 2017, the European Commission Directorate-General for Research and Innovation released the European Open Science Cloud (EOSC) Declaration, a five-page document setting out the principles of the EOSC and recommending the pursuit of a series of goals — brought together under the headings data culture and FAIR data, research data services and architecture, and governance and funding — aimed at better facilitating data driven research and its central role in the pursuit of excellent science. The EOSC will be developed as a research data infrastructure commons and the Declaration aims to secure the endorsement and commitment of all scientific stakeholders to make the EOSC a reality by 2020. The Declaration is ambitious in scope; in particular its claim that 'no disciplines, institutions or countries must be left behind' (EOSC 2017). So too is the vision of Carlos Moedas, the European Union's Commissioner for Research, Science and Innovation, for the EOSC. In his opening comments to the 'EOSC Summit' (the meeting that gave rise to the Declaration) Moedas framed the development of the EOSC as the next frontier 'of scientific collaboration and information sharing', casting it as a modern 'Republic of Letters' (2017).

Whilst the establishment of a pan-European data infrastructure is a timely and welcome initiative, the imperatives driving the foundation and vision of the EOSC may be obscuring an underlying problem which no technological development, no matter how ambitious, can independently address: current research infrastructures impede and inhibit interdisciplinarity. Moedas (2017) himself acknowledges this, stating that 'the most exciting and ground-breaking innovations are happening at the intersection of disciplines. We need to cherish and encourage this as much as we can. But right now, our current infrastructure dissuades interdisciplinary research'. The need to 'facilitate inter-disciplinarity and avoid fragmentation' or to minimise and reduce fragmentation of infrastructures is repeatedly emphasised throughout the Declaration (EOSC 2017). While reducing fragmentation may result in better conditions for the kind of interdisciplinarity the Commissioner envisions for the EOSC, the creation of a single platform alone will not achieve this. Indeed, the original 2016 report outlining the vision for the EOSC states that 'The majority of the challenges to reach a functional EOSC are social rather than technical' (Realising the European Open Science Cloud 2016).

Many of these social challenges are rooted deeply in scientific epistemic cultures. Disciplinary structures drive science in subtle and not so subtle ways, from the provision of scientific training, to the organisation of universities, to the fact that discipline-specific journals continue to dominate in particular fields in a manner that leads to targeted incentivisation and curation practices which in turn reinforce the disciplines. And with this structural disciplinarity trickling down from institutional, faculty and departmental levels to the very format we assign to our research questions which subsequently shape our ideas, comes the associated influence of what Ludwik Fleck (2012) designates as 'Denkstil', 'thought style', and 'Denkkollektiv', 'thought collective', with different thought collectives being unable to communicate with and to understand each other. The disparate scientific 'practices, social structures and infrastructures' identified in the 2016 Report as in need of a 'step change' to enable interdisciplinary knowledge creation through the EOSC are not unique to digital research infrastructures, but they have been exacerbated by the transition to Open Science and to data driven research and innovation. For example, one of the major factors now inhibiting interdisciplinary data driven research is that of how to capture provenance and facilitate the discovery, use, and reuse of discipline specific research 'data'. All data are not not created equal, and research disciplines are not all digital in the same way. Whilst Moedas (2017) sees the Cloud as a means of supporting those disciplines that may be 'lagging behind', the reality is that such disciplines may be digital in ways that do not comply with the current conceptions of what makes digital data comprehensible, valuable, and reusable. Indeed, they may resist digitisation and/or datafication. Furthermore different researchers, research disciplines and institutions, differ drastically over what they consider to be data in the first place.

This paper outlines the preliminary findings of the EU Commission's Horizon2020-funded Knowledge Complexity (KPLEX) project (https://kplex-project.eu/), which is investigating the delimiting effect digital mediation and datafication may have on rich, complex cultural data. As an ICT-35 'sister project', KPLEX is humanities-led and ICT-funded and the project partners are approaching this challenge in a comparative, multidisciplinary and multisectoral fashion. In this paper, we characterise the wider lessons to be learned from practices of data collection, access and reuse in the humanities, fields well used to dealing with complex, hybrid, and noisy data. In the next section we focus on the potential for information loss due to mismatch of criteria and guidelines for the recognition and classification of data both between and within research data infrastructures and researcher communities. The third section explores the ways in which the effectiveness of traditional signposting mechanisms such as classification systems, taxonomies, and metadata are weakened in the digital age, and set forth the implications that these forces will have for a wider set of disciplines under the vision of the EOSC. We argue that proponents of the EOSC cannot achieve the outcomes they seek if their standards are targeting a skewed

perspective of research communities' realities. The final section concludes with a brief discussion of a broader research agenda on these issues.

2. Criteria and Guidelines for the Recognition and Classification of Data: How Research Data Infrastructures can Result in Information Loss.

Traditionally libraries, data infrastructures have diversified and are now found scattered across universities, cultural heritage institutions, national repositories (such as France's HAL repository or DANS in the Netherlands), international infrastructures (like Dataverse), and project- or even person-specific web presences. These diverse infrastructures — themselves both institutions and agglomerations of resources — facilitate the sharing of data in ways that may appear or purport to be open, accessible, comprehensive, inclusive, and integrative. But this appearance of inclusivity can often be an artefact of their having, over time, acquired a perceived objectivity and/or authority that belies the curated, malleable, reactive, and performative nature not only of the infrastructures in and of themselves, but of the data they preserve. For example, the UK Arts and Humanities Data Service (which was closed down in the mid 2000s) collected the data associated with research council-funded digital projects, but not the digital interfaces created to enable interrogation of and access to these same projects. And yet these very interfaces held the key to the argument contained within the text and image files they structured and provided access to.

While the loss of useful data can be accidental or unintentional, as in the act of not archiving these digital interfaces, hindering data visibility can also be deliberate and intentional, as in the case of the human assigned keywords that facilitate access to the 55,000 video testimonies that together make up the Shoah Foundation Visual History Archive (VHA). According to the Shoah Foundation, each testimony 'average[s] a little over two hours each in length and were conducted in 62 countries and 41 languages'; with the collective testimonies totalling 'more than 115,000 hours' that have been manually indexed 'through a set of more than 65,400 keywords and phrases, 1.86 million names, and 719,000 images' (USC Shoah Foundation 2017). As discussed elsewhere by two of the authors of this paper, this manually assigned keyword index means 'it is still possible to hide materials that do not align with your usage-intention' (Edmond and Nugent-Folan 2017: 257), making it very difficult for a researcher to find data pertaining to viewpoints Shoah do not want to highlight or draw attention to, for very understandable reasons. In the case of Shoah, the omission of data is a key factor in understanding the cultural context in

which the data was created, demonstrating how the desire to make science more 'open' may ultimately hide important cultural and contextual information.

This says much about the role played by digital interfaces in terms of their capacity to both prohibit and facilitate access to data. It also highlights a further, more complex and far reaching problem that poses a major challenge to research data infrastructure and its design: any definition of data or the architecture that makes data available in an analogue or digital environment needs to maintain an awareness of the speculative potential of the information contained within its datasets. As Christine Borgman observes, 'what could be data to someone, for some purpose, at some point in time' (2015: 19) is ever changing. In the case of the UK Arts and Humanities Data Service the digital interfaces that facilitated access to data, themselves became not only data, but *lost* data, when they were not archived alongside the data they once provided access to. Data can be *anything*. It is conjectural, notional, and speculative and research infrastructures must somehow adapt and realign to this mutable status.

Knowledge production and organisation in the digital humanities can tell us a lot about what truly interdisciplinary knowledge sharing might look like, particularly when it comes to the task of recognising and classifying diverse and complex data. After all, the digital humanities is a catchall for a wide variety of research approaches, types of source material, and epistemic cultures from the highly qualitative (such as critical theory) to the highly quantitative (such as corpus linguistics). Work often revolves around the creation of digital editions (in the broadest sense of the term) of documents or collections of documents considered to have a particular aesthetic or historical value. In the course of the creation of such a digital edition, both knowledge and digital outputs are created, and use and reuse of the results *should* follow accordingly. However, this is not always the case. Findings of the 2006 LAIRAH project, led by the CIRCAh research group at UCL's Department of Information Studies, indicated that approximately 36% of the resources listed in the UK national project registry, the Humbul Humanities Hub (itself now an archived site (Webarchive.org.uk 2017)), met the criteria to be considered a 'neglected' resource (Warwick et.al. 2006: 16). While the LAIRAH team posit a number of possible reasons for this (from naming conventions to technical platforms), it is noteworthy that their primary focus was not reuse, but use. In the context of the LAIRAH project, 'use' might be defined in terms of how a resource meets its intended purpose according to the research questions and knowledge organisation frameworks. The occurrence of reuse is unquantified, but surely far lower given that,

> There may be a scholarly bifurcation between those who create specialist digital resources as part of their research, but do not tend to reuse, and those who prefer to use more generic information resources, but are less concerned with deposit and archiving (Warwick et.al. 2006: 20).

Creation and presentation of resources (including but not necessarily limited to research data) in the humanities are acts of curation, ones that are 'always already' marked by both the epistemic and organisational frameworks of the creator. This may seem a humanities-specific problem, but only when viewed superficially, as the title of Sandra Gitelman's collection of essays on data practices reminds us, *'Raw Data' is an Oxymoron* (Gitelman ed. 2013).

At a macro-level, the digital humanities and its related infrastructures also serve to remind us of the limits of data driven approaches to knowledge creation. Much of the material made available within e-research infrastructures is highly specific, relating to individual disciplines, institutions or researchers and excluding input that cannot be effectively structured, represented, or digitised. Approaches that assume the availability of all relevant primary sources in a digital format, are destined to be unsuccessful and concordantly to produce misrepresentative research (Edmond 2016). If we think about the potential of digital history to enable a richer and more accessible understanding of the past, we face a stark reminder in the results of the Enumerate survey, which shows that as of 2015, 16% of institutions surveyed had no digital collections, and only 23% of Europe's heritage was available in digital format (Enumeratedataplatform.digibis.com 2017). Given the incentives to work with openly accessible data, these numbers are striking and speak to the extent of the cultural heritage material yet to be digitised and datafied. While the percentage of cultural sources accessible in digital format and through open data archives remains so low, we must work against the presumption that the digital material available represents the totality of research material. In addition, this points to fundamental underlying issues of data in specific research environments. These environments are not 'lagging behind' the computational turn, rather they are reflecting the realities of their primary sources, and many of these realities involve complexity that prohibit successful re-presentation in a digital environment which often results in the simplification and/or misrepresentation of the primary source material.

Good infrastructure is the foundation of good science, but imperfect infrastructure can result in unintended curations that directly impinge on and influence the research that arises from it, is based upon it, and works within it. The same can be said of imperfect classification systems, which will be discussed in the next section of this paper. Data accessibility, usability, and reusability—even *what data are*—are delimited by what is provided in the metadata structures of information architecture and data infrastructures; with these infrastructures themselves performatively modifying the data they delimit. In fact, Johanna Drucker argues that metadata structures have the greatest impact on our approach to material in a digital environ:

> Arguably, few other textual forms will have greater impact on the way we read, receive, search, access, use, and engage with the primary materials of humanities studies than the metadata structures that organise and present that knowledge in digital form (Drucker 2009: 224).

This is not limited to the metadata structures that make up our information architecture. According to David Ribes and Steven Jackson, our research is increasingly influenced by 'the invisible infrastructures of data' that belie the 'occluded set of activities that produce those data themselves'. Such infrastructures include the 'Technicians, robots, and cooling systems [that] are increasingly hidden in the clouds of computing, laboring to preserve the data of the earth sciences and, agnostically, those of many others' (2013: 152). But there are other invisible infrastructures that Robes and Jackson overlook, such as the algorithms that make data, and particularly big data, available and navigable in a digital environment; a topic touched on by William Uricchio (2017: 131-32) in his account of the algorithm as 'a talisman, radiating an aura of computer-confirmed objectivity, even though the programming parameters and data construction reveal deeply human prejudices' and by Presner (2015) in his discussion of 'the ethics of the algorithm' and one the KPLEX project is exploring further.

Data with the level of complexity inherent in cultural records are perhaps more common in the humanities than elsewhere, but the technological imperative to enhance signal through the reduction of noise is everywhere, and never does it accommodate the kind of richness and potential ambiguity that most data does, on some level, contain. In an input environment where 'anything can *be data* once it is entered into a system *as data*' (Edmond and Nugent-Folan 2017: 254) data cleaning and processing, together with the metadata and information architectures that structure and facilitate our archives acquire a capacity to delimit what data are. This engenders a process of simplification that has major implications for the potential for future innovation within research environments that depend on rich material yet are increasingly mediated by digital technologies. One discipline's signal is inevitably another's noise, and all data—cultural or otherwise—are marked by the biases of the human beings that capture, clean, and curate them. This is particularly problematic when we speak, as proponents of the EOSC do, of open research data as a primary contributor to the enhancement of innovation capacity in Europe through the facilitation of increased levels and efficacy of inter- and transdisciplinary research.

3. Classification Systems, Taxonomies, Metadata.

Research data, together with the tools with which we represent them, add complexity to the relation between researchers and their objects of study. Classification systems are standardised in almost every scientific field. These classification systems represent a specific worldview and they are designed to capture and describe certain types of data. Research in Development Information Systems has

demonstrated that mismatches between ontologies of state-created information systems and local communities' representations of their contexts can lead to significant gaps between community and meta ontologies. These gaps are 'often a symptom of the fundamental difficulty of incorporating local, contextualized knowledge into large scale, comparable-across-time-and-place datasets'. The authors make it clear what is at stake in the use of distinct ontologies and describe how the desire to develop data that are comparable across communities leads to information loss, 'not just in terms of overlooked entities but more importantly in overlooked or misjudged semantic relationships between these entities' (Wallack and Srinivasan 2009). On the matter of standards, the EOSC Declaration states the following:

> The EOSC must be underpinned by minimal and rigorous global standards for open research data, as well as standards for EOSC based services for collaboration through the EOSC (e.g. to facilitate inter-disciplinarity and avoid fragmentation). These standards (technical, semantic, legal and organisational) must combine long-term sustainability with optimal freedom of local implementation. (EOSC 2017)

In the digital humanities and cultural heritage research, there are instructive examples to be found for how this kind of broad, standards-based integration and federation might or might not work, as the case may be. One of the most successful implementations of a standard in the digital humanities community has been that of the Text Encoding Initiative (TEI). Reaching back into the late 1980s, the TEI community came together with a very clear goal in mind: to address the 'overwhelming obstacle to creating sustainable and shareable archives and tools' (Tei-c.org 2017). As it exists today, the XML-based tag set described within the TEI guidelines is a known and trusted resource for preparing texts for digital representation in a flexible and interoperable way. Indeed, the success of the TEI has been such that in 2017, the entire consortium was awarded one of the digital humanities community's most prestigious awards, the Antonio Zampolli prize, for services to the research community (Tei-c.org 2017). The strength of the TEI is that, as a standard, it embodies the primary quality for an infrastructure, as described by Edwards et al in their seminal work on the topic: it gets 'below the level of the work' (Edwards et al. 2012). With the exception of a few formal elements required in the header, the TEI does not require a user to mark up particular aspects of a text, only to mark up those aspects considered important in a certain way. And the number of possibilities is vast: in its most recent release, TEI P5 contained 569 different elements to choose from (Tei-c.org 2017).

However, not every standard in the digital humanities has managed to clear the bar of being 'below the level of the work'. As described in more detail by one of the authors of this paper elsewhere, there exist a number of illustrative examples that point toward the fictionality of a model or standard, and demonstrate how mediating between digital object and a particular version of the world can obscure more than it reveals. In particular, the cases that are referred to are:

- The migration of the Europeana Digital Library from its original metadata standard (ESE) to a revised one (EDM) able to capture a sufficient level of richness about the federated objects;
- The ethical implications, good and bad, of the human cataloguing of sensitive oral histories, including their hesitation to portray victims in a negative light;
- The manner in which models for the representation of data provenance struggle to capture much of the complexity of the provenance of the historical record, where objects may pass through many hands and places in the course of being created and collected, each of which has an impact on how they might be reused or interpreted. (Edmond 2016)

The standards to be proposed for the EOSC are yet to be released to the community, but one can assume that the stated desire that they remain 'minimal' could lead them down a path of impoverished provenance, and create the risk of hiding research data from researchers whose ontologies may not match those of the data curators.

A possible indicator of how these standards might look can be found in current practice in Research Classification Systems (RCS). Research Classification Systems provide high level classification to facilitate research evaluation and quality judgements across disciplines and national systems. They classify the 'type' of research and aim to assess or provide a measure of its quality, are often developed and implemented at a National Level and are thus country specific (frequently being referred to as National Research Classification Systems). As a consequence, the metadata is not granular, tending to have a maximum of three to four facets. The practice of applying such systems can be traced back to 1963 with the release of the OECD's Frascati Manual (more formally known as *The Proposed Standard Practice for Surveys of Research and Experimental Development*) and later the Fields of Science (FOS) classifications, which came to be referred to as the *Frascati Fields of Science*. Frascati FOS provides the user or institution with a choice of classifications (which are maintained by UNESCO), and formed the basis for another major RCS, the Australia and New Zealand governments' ANZSRC (Australian and New Zealand Standard Research Classification). Like Frascati, ANZSRC is an umbrella term covering three related classifications across a wide range of disciplines: Type of Activity (TOA); Fields of Research (FOR); and Socio-economic Objective (SEO). The decision to capture data under these headings was motivated, interestingly, by the desire to create data that could and would be widely reused:

> The use of the three constituent classifications in the ANZSRC ensures that R&D statistics collected are useful to governments, educational institutions, international organisations, scientific, professional or business organisations, business enterprises, community groups and private individuals in Australia and New Zealand (Abs.gov.au 2017).

Equivalents in Europe to ANZSRS are the Common European Research Information Format (CERIF) and the European Current Research Information Systems (euroCRIS), which defines its mission as follows:

> The mission of euroCRIS is to promote cooperation within and share knowledge among the research information community and interoperability of research information through CERIF, the Common European Research Information Format. Areas of interest also cover research databases, CRIS related data like scientific datasets, (open access) institutional repositories, as well as data access and exchange mechanisms, standards and guidelines and best practice for CRIS (Eurocris.org 2017).

In spite of the ambition to transcend the commonplace of an RCS as an instrument for policy makers, and to become a living part of research, awareness and uptake of the repository and tools provided appears to have been weak, with publications on the framework peaking some years ago. Much of what is now available comes from the community that developed the standards, rather than from the researchers actively using them.

International standards for documenting and managing data have been developed for specific research communities, such as the Data Documentation Initiative (DDI), a longstanding and evolving standard 'for describing the data produced by surveys and other observational methods' for the social science research community. DDI facilitates the documentation and management of *user defined* data throughout the entirety of its lifecycle, from 'conceptualization, collection, processing, distribution, discovery, and archiving' (Ddialliance.org 2017). Within the confines of the DDI, proto-data becomes data proper simply by means of input and entry into the database in a manner that accords with the DDI metadata specifications. DDI presents data as something that is user defined, a treatment that accords with the data as an entity of speculative value, and with Raley's idea of data as performative (Raley 2013: 128). Anything can *be data* once it is entered into the system *as data*. As Borgman notes:

> The DDI is widely used in the social sciences and elsewhere for data description but does not define data per se. The DDI metadata specifications, which are expressed in XML, can be applied to whatever digital objects the DDI user considered to be data. (Borgman 2015: 20)

From this we can conclude that, like data, metadata is also performative, having the potential to situate proto-data as data proper.

Richer information can be found in some of the national systems aimed not at the registration of research activity, but its evaluation. Many such frameworks exist, with the UK's Research Excellence Framework (REF) (About - REF 2021 2017). and the Netherlands' Standard Evaluation Protocol (SEP) (Standard Evaluation Protocol 2015–2021: Protocol for Research Assessments in the Netherlands 2016). It is important to recall, however, that while these systems do capture equivalent data sets across disciplines and institutions, and do allow research activity (including the production of research data) to

be assessed for its quality (and, one might be assumed, reuse potential), their data sets can hardly be described as minimal, and barely as standardised. Different approaches are applied as appropriate between disciplines, and many aspects of the data gathering, from the UK's Impact Case Studies to the SEP's site visits, generate quite complicated, qualitative data, of the sort that can hardly be imagined for the EOSC.

Additional examples of research classification systems that might serve as an inspiration for how the EOSC might function can be found in publisher databases such as SCOPUS and Web of Science. These bibliographical databases with their abstracts and citation information, do promote discovery and reuse of research data of a sort, albeit in the more digested form of research publications in scholarly journals. It must be borne in mind, however, that the efficacy of these databases as finding aid relies upon a number of pre-existing cultural systems, including disciplinary organisation, and knowledge. Fusing the information held in such databases with the data in the EOSC could indeed be a successful and efficient starting model, upon which multidisciplinary layers could be later be built. Such a model seems relatively unlikely to emerge, however, as the corporate interests of the owners of these databases (publishing and data giants Elsevier and Thomson Reuters respectively) seem unlikely to contribute their prime assets to a public infrastructure.

If metadata standards alone are unlikely to make the EOSC realise its potential, then perhaps controlled vocabularies and classifications of other sorts hold the key? For example, the Dewey Decimal Classification (DDC) system continues to be widely used internationally and, more importantly, continues to hold a place in the scholarly imagination as the engine for the serendipitous encounter of finding the unexpected, yet ideal, book on the library shelf (Edmond et al. 2017). The Library of Congress in Washington D.C., USA, has taken over management of the Dewey system, managing it alongside their own Library of Congress Classification System (LCC). Both the DDC and LCC make use of and are driven by controlled vocabularies, as are projects such as the Getty Institute Vocabularies for art and architecture (Getty.edu 2017). Of particular interest are the controlled vocabularies adopted by these classification systems when it comes to the classification of complex data. These classification systems extend to the classification of visual data with VRA Core (which is based on Dublin Core) being a data standard for the description of visual cultural artefacts and their documenting images; VRA Core provides schemas, description and tagging protocols, and category guides. We also have the Getty Image Classification system which has subsections devoted specifically to the arts and humanities such as the AAT (Arts and Architecture Thesaurus), the ULAN (Union List of Artist Names), and CONA (Cultural Object Name Authority). These promote very definite views on classification, providing structured terminologies with the aim of making objects discoverable through standardization of classification.

What is of interest within the context of this paper in the manner in which the underpinning controlled vocabularies used or offered within these systems expose and connect diverse bodies of knowledge in a manner that is standardised, but still flexible enough to allow multiple interpretations. These promote very definite views on classification, providing structured terminologies with the aim of making objects discoverable through standardization of classification, while still enabling a multiplicity of objects, approaches and interpretations to be encompassed under their umbrella. Again, this allows for high level accessibility, but not granularity or idiosyncrasy.

These set vocabularies provide a standardised approach to the indeterminate or unknown, using words such as 'circa' for uncertain dates and terms such as 'anonymous' for uncertainty regarding authorship. This is a further example of the manner in which adaptation to the messiness of humanistic data results in accommodation between the needs for common standards and a looser hold on the precision of what is known. Alongside controlled vocabularies governed by set thesauri, there are also locally used classifications and folksonomies, which feature locally defined classification systems or, in the case of folksonomies, 'bottom-up' user contributed keywords or user-generated tagging mechanisms. Folksonomies themselves pose further problems, particularly in relation to the risk of introducing further ambiguity with 'ambiguous headings' having been identified in one Canadian study as 'the most problematic area in the construction of the tags; these headings take the form of homographs and abbreviations or acronyms' (Spiteri 2007). Drawing on the work of Quintarelli and Fichter, Louise Spiteri notes that folksonomies 'reflect the movement of people away from authoritative, hierarchical taxonomic schemes; the latter reflect an external viewpoint and order that may not necessarily reflect users' ways of thinking' (Spiteri 2007).

The above survey provides a matrix of metadata, controlled vocabulary, taxonomies, and classification systems that make up the cataloguing and metadata rules that have been adopted by Galleries, Libraries, Archives and Museum (GLAM) institutions. Whether the EOSC architecture will be able to accommodate this diverse mix of standards and descriptors to facilitate the interdisciplinarity it is mandated to promote remains to be seen.

4. Conclusion.

We are not proposing the reconstitution of research data management systems. Rather, we need to replace the tendency to skeuomorph by modelling our digital catalogues on library catalogues (which are generally accompanied by the gentle hand and deep knowledge of the librarian behind them) and instead

develop models that better capture the lines not just between scholarly disciplines, but between the analogue and the digital, between formal and informal knowledge, between the search engine and the knowledgescape. It is impossible to tell where the EOSC will take us in the next two years, much less in the next ten. The experiences of the humanities imply, however, that if the resource is truly to fulfill its stated aims and ambitions, a balance will need to be struck between an easy to create 'minimal' description, which may become a barrier to reuse, and a rich description that is more granular, but this granularity concordantly becomes a barrier to deposit. A balance must be struck between standards for longevity, sustainability, and interoperability, and the facilitation of serendipity, discoverability, and comprehensibility across epistemic lines. This paper employs anecdotes from the humanities to identify a number of potentially significant challenges for the development of the EOSC. These challenges come in the form of extant practices of data collection, access, and reuse and in particular the problems associated with classifying complex data. Similarly, the potential for information loss between and within research data infrastructures and researcher communities is not insignificant and poses a major challenge to the pan-disciplinary vision of EOSC; one that entails a reassessment of the effectiveness of traditional signposting mechanisms such as classification systems, taxonomies, and metadata are weakened in the digital age. Should we be able to realise and reconcile these challenges, the vision of the EOSC as a platform that leaves no disciplines, institutions or countries behind will have the strongest possible chance of realisation.

Acknowledgements

This work was developed in the context of the project Knowledge Complexity (KPLEX). The KPLEX project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 732340.

References

Abs.gov.au, (2017). *1297.0 - Australian and New Zealand Standard Research Classification (ANZSRC), 2008*. [online] Available at: http://www.abs.gov.au/Ausstats/abs@.nsf/Latestproducts/1297.0Main%20Features32008 [Accessed 17 Nov. 2017].

Ddialliance.org, (2017). *Welcome to the Data Documentation Initiative | Data Documentation Initiative*. [online] Available at: https://www.ddialliance.org/ [Accessed 17 Nov. 2017].

Drucker, J. (2009). *SpecLab: Digital Aesthetics and Projects in Speculative Computing*. Chicago and London: University of Chicago Press.

Edmond J. (2016). Will Historians Ever Have Big Data?. In: B. Bozic, G. Mendel-Gleason, C. Debruyne, D. O'Sullivan, eds, *Computational History and Data-Driven Humanities*. CHDDH 2016. IFIP Advances in Information and Communication Technology, vol 482. Springer, Cham. Doi: 10.1007/978-3-319-46224-0_9.

Edmond, J., Bagalkot, N., and O'Connor, A. (2017). Toward a Deeper Understanding of the Scientific Method of the Humanist. [online] Available at: https://hal.archives-ouvertes.fr/hal-01566290. [Accessed 17 Nov. 2017].

Edmond, J. and Nugent-Folan, G. (2017). Data, Metadata, Narrative. Barriers to the Reuse of Cultural Sources. *Communications in Computer and Information Science*, 755, pp. 253-260.

Edwards, P., Jackson, S., Bowker, G., and Knobel, C. (2012). Understanding Infrastructure: Dynamics, Tensions and Design. [online] Available at http://hdl.handle.net/2027.42/49353 [Accessed 17 Nov. 2017].

Enumeratedataplatform.digibis.com. (2017). *Survey Report on Digitisation in European Cultural Heritage Institutions 2015 - ENUMERATE Data Platform*. [online] Available at: http://enumeratedataplatform.digibis.com/reports/survey-report-on-digitisation-in-european-cultural-heritage-institutions-2015/detail [Accessed 16 Nov. 2017].

Eurocris.org.(2017). *What is euroCRIS? | euroCRIS*. [online] Available at: http://www.eurocris.org/what-eurocris [Accessed 17 Nov. 2017].

European Open Science Cloud (EOSC) (2017). *EOSC Declaration*. [online] Available at: https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud [Accessed 16 Nov. 2017].

Fleck, L. (2012). *Genesis and Development of a Scientific Fact*. Chicago and London: University of Chicago Press.

Getty.edu. (2017). *Getty Vocabularies Editorial Guidelines (Getty Research Institute)*. [online] Available at: http://www.getty.edu/research/tools/vocabularies/guidelines/index.html [Accessed 17 Nov. 2017].

Gitelman, L. ed., (2013). *'Raw Data' is an Oxymoron*. Massachusetts: MIT Press.

Moedas, C. (12 June 2017). EOSC Summit: The European Open Science Cloud – The New Republic of Letters.

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

Raley, R. (2013). Dataveillance and countervailance. In: L. Gitelman, ed., *'Raw Data' is an Oxymoron*, 1st ed. Massachusetts: MIT Press, pp. 121-146.

Realising the European Open Science Cloud: First report and recommendations of the Commission High Level Expert Group on the European Open Science Cloud, (2016). [ebook] Brussels: European

Commission. Available at: https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf#view=fit&pagemode=none. [Accessed 17 Nov. 2017].

Ref.ac.uk, (2017). *About - REF 2021*. [online] Available at: http://www.ref.ac.uk/about/ [Accessed 17 Nov. 2017].

Ribes, D. and Jackson, S. (2013). Data bite man: The work of sustaining a long-term study. In: L. Gitelman, ed., *'Raw Data' is an Oxymoron*, 1st ed. Massachusetts: MIT Press, pp. 147-166.

Standard Evaluation Protocol 2015 – 2021: Protocol for Research Assessments in the Netherlands. (2016). 3rd ed. [ebook] The Netherlands: Association of Universities in the Netherlands (VSNU), Netherlands Organisation for Scientific Research (NWO), the Royal Netherlands Academy of Arts and Sciences (KNAW). Available at: http://www.vsnu.nl/files/documenten/Domeinen/Onderzoek/SEP2015-2021.pdf [Accessed 17 Nov. 2017].

Spiteri, L. F. (2007). *Structure and form of folksonomy tags: The road to the public library catalogue*. [online] Available at: http://www.webology.org/2007/v4n2/a41.html [Accessed 30 Nov. 2017].

Tei-c.org. (2017). *Appendix C Elements - The TEI Guidelines*. [online] Available at: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/REF-ELEMENTS.html [Accessed 30 Nov. 2017].

Tei-c.org. (2017). *TEI: History*. [online] Available at: http://www.tei-c.org/About/history.xml; http://www.tei-c.org/index.xml [Accessed 16 Nov. 2017].

Uricchio, W. (2017). Data, culture and the ambivalence of algorithms. In M.T. Schäfer, and K. van Es, eds, *The Datafied Society. Studying Culture through Data*. Amsterdam: Amsterdam University Press, pp. 125–138.

USC Shoah Foundation. (2017). *About Us*. [online] Available at: https://sfi.usc.edu/about [Accessed 16 Nov. 2017].

Wallack, J. and Srinivasan R. (2009). Local-global: Reconciling mismatched ontologies in development information systems. In: *Proceeding of the 42nd Hawaii International Conference on System Sciences*, 2009. Washington, DC: IEEE Computer Society. DOI: 10.1109/HICSS.2009.295.

Warwick, C., Terras, M., Huntington, P., Pappa, N., and Galina, I. (2006). The LAIRAH project: log analysis of digital resources in the arts and humanities. Final report to the Arts and Humanities Research Council. Project Report. [online] Swindon: Arts and Humanities Research Council. Available at: http://dro.dur.ac.uk/15196/1/15196.pdf [Accessed 16 Nov. 2017].

Webarchive.org.uk. (2017). *UK Web Archive*. [online] Available at: http://www.webarchive.org.uk/ukwa/target/125037/source/alpha [Accessed 16 Nov. 2017].