# Seeing Shapes in the Cloud: Perspectives from the Humanities on Interdisciplinary Data Integration

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 Postdoctoral Research Fellow and Project Officer at the Centre for Digital Humanities, Trinity College Dublin, Ireland and is presently contributing to the *PROgressive VIsual DEcision-Making in Digital Humanities* (PROVIDEDH) project (https://providedh.eu/). She previously worked as a Research Assistant on the Horizon 2020 Knowledge Complexity (KPLEX) project (www.kplex-project.eu). She is 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 Associate Professor of Digital Humanities 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 President of the Board of Directors of the DARIAH ERIC, a body she represents on the European Commission's Open Science Policy Platform (OSPP). She was Principal Investigator on the Horizon 2020 Knowledge Complexity (KPLEX) project (www.kplex-project.eu).

**Georgina Nugent-Folan** is Assistant Professor of Modern English Literature at LMU Munich, Germany. She previously worked as a Postdoctoral Research Fellow at the Centre for Digital Humanities, Trinity College Dublin, Ireland where she contributed 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 an approach based upon the findings of the Knowledge Complexity (KPLEX) project (https://kplex-project.eu/) which investigated the delimiting effect digital mediation and datafication can have on rich, complex cultural data. As KPLEX was humanities-led but ICT-funded, the project partners approached 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 arts and 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 key policy-relevant findings on how to tackle those issues and maximise the participation of 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,[1] 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 in Europe and recognising its central role in the pursuit of excellent science. The EOSC would be developed as a research data infrastructure commons and the Declaration aimed to secure the endorsement and commitment of all scientific stakeholders to make the EOSC a reality by 2020. This aim has since been progressed through the initial launch of an EOSC portal in November 2018, and continues to develop as an ecosystem of data, services and tools. The Declaration was ambitious in scope, in particular in its claim that 'no disciplines, institutions or countries must be left behind' (EOSC 2017) as was the vision for it promoted by Carlos Moedas, the European Union's then Commissioner for Research, Science and Innovation. 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, vision and ongoing development of the EOSC may be obscuring an underlying problem which no technological development, no matter how ambitious, can independently address: namely that research infrastructures are generally designed to serve communities of specialists, or to preserve cognate collections of research objects, and therefore are not optimised to facilitate interdisciplinarity. Moedas (2017) himself acknowledged 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).

[1] A Glossary of Acronyms has been provided at the end of this paper.

While reducing fragmentation may result in better conditions for the kind of interdisciplinarity the Commissioner envisioned 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 stated that 'The majority of the challenges to reach a functional EOSC are social rather than technical' (European Commission 2016: 6).

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 the research questions which subsequently shape our ideas, there comes an associated influence of what Ludwik Fleck (2012) designated as 'Denkstil', 'thought style', and 'Denkkollektiv', 'thought collectives', 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 (European Commission 2016: 9), but they have been exacerbated by the transition to Open Science and to data-driven research and innovation. All data are not not created equal, and research disciplines are not all digital in the same way.

For example, two 2017 surveys of the EU language technology community found that access to and availability of language resources and language processing tools for many EU languages other than English and the FIGS languages (French, Italian, German, and Spanish), continues to be a major issue. The most frequently encountered issues with language resources are data availability in the first instance, followed by openness of data, and Intellectual Property Rights (IPR) issues. Language technology solutions are developed using language data as input material, therefore data issues — such as errors, noise, and inconsistencies in coverage — have a crucial impact on the quality of solutions. These issues are often exacerbated by data scarcity and data inequality, particularly for smaller languages and overlooked domains (Kalnins and Vasiljevs 2018). Whilst Moedas (2017) saw the Cloud as a means of supporting those disciplines that might be 'lagging behind', the reality is that such disciplines may not be lagging behind at all, but rather they may be digital in ways that do not comply with other, dominant conceptions of what makes digital data comprehensible, valuable and reusable. Indeed, they may resist

digitisation and/or datafication. Furthermore, individual researchers, research disciplines and institutions, differ drastically over what they consider to be data in the first place.

This paper approaches the problems outlined above with specific reference to the arts and humanities, two of the most prominent of those potentially 'lagging' communities. In particular, it looks at some of the key elements of common datification and data reuse practices, such as will underscore the EOSC, and the material and knowledge-based reasons why these specific disciplines may resist them. Considering that the data informing research in the arts and humanities is different only in degree, and not in kind, from that used by other disciplines, our exposition of these points of friction and resistance will provide insights useful far beyond the fields underlying this enquiry. By focussing on the sources of lag in the disciplines operating largely in the so-called 'long tail' of data-driven research (Doorn 2018), we identify a number of potentially significant challenges for the development of the multidisciplinary data infrastructure envisioned for the EOSC. We argue that proponents of the EOSC, or indeed of any ambitious multi-disciplinary data integration project, will not be able to achieve the outcomes they seek if the resulting infrastructures are based on a skewed perspective of research communities' realities. To conclude, we also offer a brief discussion of a broader research agenda on these issues.

# 2. Methodology.

This paper draws upon the findings of the Horizon 2020-funded Knowledge Complexity (KPLEX) project (https://kplex-project.eu/). The core mission of the KPLEX project was to define and describe the key aspects of rich, complex cultural data that are at risk of being left out of our knowledge creation processes in a system where large scale data aggregation is becoming ever more accepted as the gold standard. As a so-called 'sister project', intended to build dialogue between diverse research domains, the KPLEX project partners approached these questions via a humanities research perspective, but using a variety of research tools. The various methods employed by the project team included a large and multiperspectival information gathering exercise, covering four distinct surveys (completed by over 700 respondents) and 38 expert interviews which were subsequently coded and analysed.[2] The groups from which participants were drawn included a range of perspectives from around the practice of big data research, including cultural heritage practitioners, researchers from humanities, science and engineering, and technology developers.

[2] The KPLEX research dataset, including redacted interview transcripts are available here: https://doi.org/10.17026/dans-xe6-hpm5.

The challenges identified in this paper are inspired by a specific strand of research within the project titled *Toward a New Conceptualisation of Data* (Edmond and Nugent Folan 2018). The aim of this work was to examine, among other things, conceptualisations of the term 'data' and its many possible meanings and implications as enacted in the humanities and in computer science. Furthermore, this research sought to understand and to articulate the limitations that simplified conceptualisations of data may place on innovation. The primary methods employed for this investigation were an extensive literature review of the state of knowledge regarding the concept of data, combined with empirical observations of humanistic and computer science research and development. The empirical enquiry into leading researchers' practices and narratives regarding cross-disciplinary collaboration in terms of methodologies, datafication, data management and sharing illustrates basic epistemological challenges to data aggregation in the context of heterogeneous data sets.

# 3. How Data Management can Result in Information Loss.

Although traditionally limited to functioning within libraries and archives, data infrastructures have diversified and are now found scattered across universities, cultural heritage institutions, national repositories, international infrastructures, 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 more open, accessible, comprehensive, inclusive, and integrative than their analogue counterparts. But this appearance can often be an artefact of their having, over time, acquired a perceived objectivity and/or an 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 2007) 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 (ahds.ac.uk 2017). And yet these very interfaces often hold the key to the argument contained within the text and image files they structure and make accessible. Form is as much a part of the argument of a work of scholarship as content, an authority that has been passed from the carefully formulated language of an article to the intricacies of the interface or data organisation framework of a digital edition or database.

While such a loss of useful context (and ultimately data) regarding the preparation, organisation and curation of a dataset can be accidental or unintentional — as in the act of not archiving these digital interfaces — hindering data visibility can also be deliberate and intentional. An illustrative case of this is that 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 USC 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' (Shoah Foundation 2017). The Shoah VHA team adopted a standardised vocabulary through their adherence to the NISO Z39.19 standard. The goal of NISO Z39.19 is to provide 'guidelines and conventions for the contents, display, construction, testing, maintenance, and management of monolingual controlled vocabularies' (NISO 2010). The keyword principles employed in the Shoah VHA 'derive from the application of a specific standard (Z39.19) to consistently and unambiguously describe "content objects" (the survivor testimonies) in order to produce a monolingual controlled vocabulary (the thesaurus) to facilitate their search and retrieval'. Hierarchical vocabularies and taxonomies form the structure of search programmes such as that used by the Shoah VHA, and these are tiered so that 'under each of these broad categories are hierarchical vocabularies to facilitate searching at a more precise level' (Presner 2015). The standard was applied with the intention of creating an objective and robust structure for understanding these testimonies, but that may not have been the final effect. In his analysis of the data practices of the VHA, Presner notes that:

> The question of what constitutes a keyword is the starting point for query design, for that is what makes querying and query design practically part of a research strategy. When formulating a query, one often begins with keywords so as to ascertain who is using them, in which contexts and with which spread or distribution over time (2015).

It is therefore arguable that the question of what constitutes a keyword is also the starting point for epistemic and ethical queries, and the Shoah VHA keyword index is problematic from this perspective. This is the *only* way to search through the archive in its totality, so this in itself is problematic because it limits searchability to the keywords chosen manually by human project contributors. Presner argues that with respect to the Shoah VHA: 'we see a symbiosis between narrative and database, such that the paradigmatic structure of the database contributes to the syntagmatic possibilities of combination at the heart of narrative' (Presner, 2015). Subjective selection for the purpose of narrative formation takes place at a structural level within this archive, and the 'syntagmatic possibilities' are limited to those imposed by the people who assigned the keywords.

This example says much about the role played by digital interfaces in terms of their capacity to both prohibit and facilitate access to complex data, as most, if not all, data held by cultural institutions tends to be. When we speak of complex data we are referring to the fact that in the process of digitisation,

much valuable information is often lost: ‘We generally capture only part of a phenomenon, thereby reducing its complexity for the next user of that surrogate’ (Nugent Folan 2018). The records of human activity and creativity we refer to here as cultural data cling to the complexity of their contexts for a number of reasons particularly pronounced, albeit not unique, within the ecosystem of research data. This is due to a number of factors, including the gap between the latent and manifest purposes behind their creation, the temporal gap between the time of their creation and their digitisation, their multidimensionality as material objects, and their tendency to accrete allied layers of meaning through the institutions that harbour them, the provenance that shapes understanding of them and tacit knowledge of professionals that take on responsibility for them.

The example of the Shoah VHA also highlights a further, more complex and far reaching problem that poses a major challenge to broadly interdisciplinary 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. Similarly, Sabina Leonelli frames the notion of data as ‘a relational category applied to research outputs that are taken, at specific moments of inquiry, to provide evidence for knowledge claims of interest to the researchers involved’ (2015). Data can be *anything*. It is conjectural, notional, and speculative and research infrastructures must somehow adapt and realign to this mutable status. The future of data-driven research will lose greatly in its richness if the current conceptions of data continue to be so divergent, with some disciplines using the word to mean a wide variety of things, from input to output, machine-produced to human produced, simple strings to complex objects, and others seeing it as an irrelevant term for their work (Edmond and Nugent Folan 2019). If data is to be understood as a fundamental, basic building block of interdisciplinary enquiry, enabling knowledge creation across the differing axes of complexity and context instilled by the epistemic cultures that created them, much work will need to be done to develop greater transparency around this term among disciplines currently defining it, and its possible usage, requirements and limitations, very differently.

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 an umbrella term 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 composite digital objects, 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 Log Analysis of Internet Resources in the Arts and Humanities (LAIRAH) project, which aimed to discover the influences of long-term sustainability and use of digital resources in the humanities, 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 posited 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, which can take a project results (and in particular its data) beyond the mode of interrogation imagined by its creator, 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) are always, and particularly in the humanities, acts of curation. Data 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 Lisa Gitelman's collection of essays on data practices reminds us, *'Raw Data' is an Oxymoron* (Gitelman 2013).

At a macro-level, the digital humanities and its related e-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. The enduring dominance of research questions framed outside of a data-driven paradigm mean that approaches assuming the availability of all relevant primary sources in a digital format are generally 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 (or, indeed, likely never to be) digitised and datafied. While the percentage of cultural sources accessible in digital format and through open data archives remains so low, humanists in particular must work against the presumption that the digital material available represents the totality of research material, though this is a fallacy that could in theory impact upon any discipline. In addition, this points to fundamental underlying issues of data in specific research contexts. These bodies of research are not 'lagging behind' the computational turn, rather they are reflecting the realities of their primary sources and external referents, and many of these realities involve a complexity that prohibits successful re-presentation in digital or datafied environments as they have currently been conceived, as to adapt content to these environments can result in the simplification and/or misrepresentation of the primary source material. A medieval manuscript can be imaged and transcribed and shared as data, but regardless of image quality, still much is lost: do we need imaging to the chemical level, so that the origin of the ink can be determined? Or to the genetic level, to preserve information on the origin of the hide it has been written on?

Good infrastructure is the foundation of good science, but imperfect infrastructure can result in the foundation of knowledge upon unrecognised biases that directly impinge on 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* — these concepts 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 environment:

> 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).

Such arguments are 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 Ribes and Jackson do not detail, 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.'

Nevertheless, 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.

# 4. A Brief Survey of Digital Humanities and Cultural Heritage Data Standards.

All data entail certain metadata resulting from the manner in which they are created, managed and used, and classification systems are standardised in almost every scientific field. That these classification systems represent a specific worldview and that they are designed to capture and describe certain types of data is well documented by Bowker and Star (2000). On the matter of standards, the EOSC Hub Technical Architecture and Standards Rodamap, which is currently under review, states that: "Metadata must be provided in a standardised format and schema and made available and accessible for harvest requests and some mandatory fields (e.g a title and data identifier) must be provided. In the next stage, refinement and enrichment of the metadata is done iteratively" (EOSC-hub 2020). Cultural heritage information professionals, including Digital Humanities researchers, often extend the meaning of the term metadata to the additional information they create to catalogue, index and otherwise enhance access to data (Gilliland 2016). In the remainder of this paper we introduce the types of data standards as defined and employed by these communities, as this provides a mature and sound body of work for how the kind

of broad, standards-based integration and federation envisioned for the EOSC might or might not work, as the case may be.

### *4.1 Data Standards in the Digital Humanities*

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' (2007: 17). 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 metadata standard in the Digital Humanities has managed to clear the bar of being 'below the level of the work'. There exist a number of illustrative examples that point toward the fictionality of any model or standard, and the fact that their authority must be earned through community engagement and involvement rather than imposed or assumed. Such examples demonstrate how mediating between digital objects and a particular version of the world can obscure more than it reveals. The case of the Shoah VHA, and the ethical implications, good and bad, of the human cataloguing of sensitive oral histories, including their hesitation to portray victims in a negative light, has already been discussed above. But others include:

- the failure of the Europeana Digital Library's original metadata standard, the Europeana Standard Elements (ESE), to capture enough information to create an effective bridge between contributing institutions and users, requiring migration to a revised Europeana Data Model (EDM) able to capture a sufficient level of richness about the federated objects (Doerr et al. 2010);
- the manner in which models for the representation of data provenance, such as the W3C iProv, are seen by collection experts as inadequate to capture the complexity of the provenance of a

historical record or object, which 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);

- and many standards, such as the W3C EmotionML Markup Language, which is not widely used by researchers because it is not viewed as encompassing an appropriate level of complexity, and isn't easy to apply (Huber et al. 2018).

### *4.2 International Data Management Standards Developed by Individual Research Communities*

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). 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. The success of DDI, much like that of the TEI discussed above, therefore sounds a warning bell regarding the size and nature of a community that can build and maintain a meaningful standard for shared research objects.

### *4.3 Research Classification Schemes*

A possible indicator of how these standards might look can be found in current practice in Research Classification Schemes (RCS). RCS 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, and are often developed and implemented at a National Level thus making them generally 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 forms 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 development on and discussions of the framework no longer very active. Much of what is now available comes from the community that developed the standards, rather than from the researchers actively using them.

Additional examples of RCS 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.

### *4.4 National Research Evaluation Systems*

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, such as 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 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, or as an easy entry portal into interdisciplinary reuse.

### *4.5 Controlled Vocabularies*

If data standards or RCS alone are unlikely to facilitate widespread sharing of data and disciplinary knowledge, then perhaps additional 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 is the manner in which the underlying 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. They promote very definite views on classification, however, 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 research data results in accommodation between the needs for common standards and a looser hold on the precision of what is known, such as the human brain is easily able to assimilate. 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 uncertainty 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 Emanuelle Quintarelli (2005) and Darlene Fichter (2006), Louise Spiteri (2007) 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'.

*4.6 Implications for EOSC Technical Architecture and Standards*

The process of transforming culture into data necessitates the classification of its various elements into taxonomies and ontologies. Research by Wallack and Srinivasan (2009) in Development Information Systems has demonstrated that mismatches between distinct ontologies (that is, 'distinction[s] between groups' mental maps of their surroundings') 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'. Wallack and Srinivasan continue by stating what is at stake in the use of specific 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: 2). Gilliland (2016) draws a similar conclusion about attempts within and between individual members of the GLAM community to integrate related materials when she notes that:

> the distinctiveness of the various professional and object-based approaches (e.g., widely differing notions of provenance and collectivity as well as of structure), different institutional cultures, and divergent cultural approaches (e.g., those exemplified in indigenous protocols for archival and library materials) have left many professionals, and the communities they represent, feeling that their practices and needs have been shoehorned into structures that were developed by another community with quite different epistemologies, practices, and users.

The data standards to be proposed for the EOSC are yet to be released to the community: even with the portal now released inclusion is by application and the only defined characteristics are that data adhere to the FAIR principles (eosc-portal.eu 2020), 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, be they formal or informal, may not match those of the data curators.

The above survey provides a matrix of metadata standards and schema, controlled vocabulary, taxonomies, and classification systems that make up the cataloguing and metadata rules that have been adopted by cultural heritage information professionals and 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.

# 5. Conclusion.

It is not the purpose of this paper to propose the reconstitution of research data management systems. Rather, if we are to reach the goal of data clouds that can provide access to a wide range of research objects and support widespread interdisciplinary convergence, we need to overcome the tendency towards skeuomorphism, which leads us to model our digital catalogues on library catalogues (which are generally accompanied by the gentle hand and deep knowledge of the librarian behind them). Instead, the key to this promised land surely lies in developing 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. Effective convergence at community level needs to be extended, but not by standardisation efforts that reduce differences between epistemic and data cultures, but which provides flexible and context-rich bridges between them. All disciplines must be encouraged to see the richness of the data layers in their work, and all researchers must be incentivised to share, for preparing data for reuse is laborious, and does not have an immediate return for the scientists undertaking such work. An expanded role for research libraries should be considered in this respect, as well as for workflows that harness, rather than replicate or come into conflict with, existing research processes.

It is impossible to tell where the EOSC will take us. 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, without this granularity concordantly becoming 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. The challenges inherent in these competing needs will challenge 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.

# Glossary of Acronyms

| ITEM | DESCRIPTION |
|---|---|
| AAT | Arts and Architecture Thesaurus |
| ANZSRC | Australian and New Zealand Standard Research Classification |
| CERIF | Common European Research Information Format |
| CONA | Cultural Object Name Authority |
| DDC | Dewey Decimal Classification |
| DDI | Data Documentation Initiative |
| EDM | Europeana Data Model |
| EOSC | European Open Science Cloud |
| ESE | Europeana Standard Elements |
| euroCRIS | European Current Research Information Systems |
| FAIR | Findable, Accessible, Interoperable, Reusable |
| FOR | Fields of Research |
| FOS | Fields of Science |
| GLAM | Galleries, Libraries, Archives and Museum |
| KPLEX | Knowledge Complexity Project |
| LARIAH | Log Analysis of Internet Resources in the Arts and Humanities |
| LCC | Library of Congress Classification System |
| NISO | National Information Standards Organization |

| | |
|---|---|
| OECD | Organisation for Economic Co-operation and Development |
| RCS | Research Classification Systems |
| REF | Research Excellence Framework |
| SEO | Socio-economic Objective |
| SEP | Standard Evaluation Protocol |
| TEI | Text Encoding Initiative |
| TOA | Type of Activity |
| ULAN | Union List of Artist Name |
| UNESCO | United Nations Educational, Scientific and Cultural Organization |
| VHA | Visual History Archive |
| VRA | Visual Resources Association |
| W3C | World Wide Web Consortium |
| XML | Extensible Markup Language |

# References


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

**Ahds.ac.uk** 2017. Available at: http://www.ahds.ac.uk/ [Last accessed 25 Sep. 2020].

**Borgman, C** 2015 *Big Data, Little Data, No Data*. Cambridge, Massachusetts: MIT Press.

**Bowker, G** and **Star, S** 2000 *Sorting Things Out: Classification and its Consequences*. Massachusetts: MIT Press

**Ddialliance.org** 2017 *Welcome to the Data Documentation Initiative | Data Documentation Initiative*. Available at: https://www.ddialliance.org/ [Last accessed 17 Nov. 2017].

**Doerr, M**, **Gradmann, S**, **Hennicke, S**, **Isaac, A**, **Meghini, C**, and **Sompel, H** 2010 The Europeana Data Model (EDM). In: World Library and Information Congress: 76th IFLA General Conference and Assembly. pp. 10-15.

**Doorn, P** 2018 *The Long Tail of Data Science: A PLAN-E Workshop in the Context of the EOSC. Workshop Report PLAN-E Plenary Paris, 19-20 April 2018*. Available at: https://planeurope.files.wordpress.com/2018/10/report-plan-e-workshop-the-long-tail-of-science-and-data-version-1-0.pdf [Last accessed 24 Sep. 2020].

**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.

**Edmond, J**, **Bagalkot, N**, and **O'Connor, A** 2017 Toward a Deeper Understanding of the Scientific Method of the Humanist. Available at: https://hal.archives-ouvertes.fr/hal-01566290. [Last 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: 253-260.

**Edmond, J** and **Nugent Folan, G** 2018 D2.1 Redefining what data is and the terms we use to speak of it. Available at: https://kplexproject.files.wordpress.com/2018/07/d2-1-redefining-what-data-is-and-the-terms-we-use-to-speak-of-it.pdf. [Last accessed 25 Sep 2020].

**Edmond, J** and **Nugent Folan, G** 2020 Digitising Cultural Complexity: Representing Rich Cultural Data in a Big Data Environment. In: Rice, R., Yates, S. (eds) *The Oxford Handbook of Digital Technology and Society*. Oxford, UP.

**Edmond, J** 2018 (Trinity College Dublin): Knowledge Complexity. DANS. https://doi.org/10.17026/dans-xe6-hpm5.

**Edwards, P**, J**ackson, S**, **Bowker, G**, and **Knobel, C** 2007 Understanding Infrastructure: Dynamics, Tensions and Design: Report of a Workshop on "History & Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures. Available at http://hdl.handle.net/2027.42/49353 [Last accessed 17 Nov. 2017].

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

**Eurocris.org** 2017 *What is euroCRIS? | euroCRIS*. Available at: http://www.eurocris.org/what-eurocris [Last accessed 17 Nov. 2017].

**European Commission** 2016 *Realising the European Open Science Cloud: First report and recommendations of the Commission High Level Expert Group on the European Open Science Cloud*. [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. [Last accessed 17 Nov. 2017].

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

**European Open Science Cloud Portal** 2020 *For Partners*. Available at: https://eosc-portal.eu/for-providers [Last accessed 04 September 2020].

**EOSC-Hub** 2020 D10.4 EOSC Hub Technical Architecture and standards roadmap v2. Available at: https://www.eosc-hub.eu/deliverable/d104-eosc-hub-technical-architecture-and-standards-roadmap [Last accessed 22 Sep. 2020].

**Fichter, D** 2006 Intranet applications for tagging and folksonomies. *Online* 30(3): 43-45

**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)*. Available at: http://www.getty.edu/research/tools/vocabularies/guidelines/index.html [Last accessed 17 Nov. 2017].

**Gililand, A** 2016 Setting the Stage. In: M. Baca, ed., Introduction to Metadata, 3rd ed. Los Angeles: Getty Research Institute. Available at: https://www.getty.edu/publications/intrometadata/setting-the-stage/ [accessed 20 April 2021].

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

**Huber, E, Lehmann, J, Stodulka, T**, 2018 D4.1 KPLEX – Report on Data, Knowledge Organisation and Epistemics – 2018-03-30. Available at: https://kplexproject.files.wordpress.com/2018/06/k-plex_wp4_report-data-knowledge-organisation-epistemics.pdf. [Last accessed 30 Nov. 2018].

**Kalnins, R** and **Vasiljevs, A** 2018 D5.1 Report on Multilingual Big Data and Language Technology. Available at: https://kplexproject.files.wordpress.com/2018/06/kplex_wp5-deliverable.pdf. [Last accessed 20 April 2021].

**Leonelli, S** 2015 What Counts as Scientific Data? A Relational Framework. *Philosophy of Science*, 82: 810–821.

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

**NISO** 2010 ANSI/NISO Z39.19-2005 (R2010): Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies. Available at: https://groups.niso.org/apps/group_public/download.php/12591/z39-19-2005r2010.pdf [accessed 25 Sep. 2020].

**Nugent Folan, G** 2018 *A KPLEX Primer for the Digital Humanities*. Available at: https://kplex-project.com/a-kplex-primer-for-the-digital-humanities/ [Last accessed 25 Sep. 2020].

**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.

**Quintarelli, E** 2005 Folksonomies: Power to the people. Available at: http://www.iskoi.org/doc/folksonomies.htm [Last accessed 25 Sep. 2020].

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

**Ref.ac.uk** 2017 *About - REF 2021*. Available at: http://www.ref.ac.uk/about/ [Last 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 [Last accessed 17 Nov. 2017].

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

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

**Tei-c.org** 2017 *TEI: History*. Available at: http://www.tei-c.org/About/history.xml; http://www.tei-c.org/index.xml [Last 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*. Available at: https://sfi.usc.edu/about [Last 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. Swindon: Arts and Humanities Research Council. Available at: http://dro.dur.ac.uk/15196/1/15196.pdf [Last accessed 16 Nov. 2017].

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