What Can a Knowledge Complexity Approach Reveal About Big Data and Archival Practice?

Nicola Horsley
DANS
KNAW
The Hague, The Netherlands
nicola.horsley@dans.knaw.nl

Abstract— **As one of the major technological concepts driving ICT development today, big data has been touted as offering new forms of analysis of research data. Its application has reached out across disciplines but some research sources and archival practices do not sit comfortably within the computational turn and this has sparked concerns that cultural heritage collections that cannot be structured, represented, or, indeed, digitised accordingly may be excluded and marginalised by this new paradigm. This work-in-progress paper reports on the contribution of the KPLEX project's knowledge complexity approach to understanding the relationship between big data and archival practice.**

Keywords- big data; knowledge compexity; digital humanities

I. Knowledge Complexity

KPLEX is investigating issues of complexity in humanities knowledge to further understanding of how the rich data that researchers of cultural heritage are so adept at creating knowledge from might be at risk in the turn towards approaches that rely upon the interrogation of large data corpora. Our research aims to illuminate the gap between analogue or augmented digital practices and fully computational ones by focusing on three key challenges to the knowledge creation capacity of big data approaches: the manner in which data that are not digitised or shared become 'hidden' from aggregation systems; the fact that data are created by humans, and lack the objectivity often ascribed to the term; and the subtle ways in which data that are complex almost always become simplified before they can be aggregated. KPLEX seeks to define and describe some of the key aspects of data that are at risk of being left out of our knowledge creation processes when large scale data aggregation is becoming ever more accepted as the gold standard of research.

This paper will describe how our approach to knowledge complexity elucidates the concept of 'hidden data' in particular. We are using the term 'hidden' to simply denote data that are at risk of not being used. This may be because potential users cannot access data because they have not been digitised, cannot find them because metadata have not been shared or because of cultural gaps between data use and archival practice such as the general shift in expectations towards instantaneous information retrieval, with barriers to data access (which have always existed) perceived as confirmation that the desired data simply do not exist.

Our knowledge complexity approach recognises the efforts of developers of platforms like Europeana (www.europeana.eu/portal/en) to promote the discoverability of data in cultural heritage institutions, while acknowledging that many more institutions hold back from this kind of computational thinking. Indeed, in engaging with cultural heritage practitioners from a range of institutions, we anticipate resistance to the computational turn to be born out of diverse practices and principles. Archival institutions have a long history of professional development to protect their holdings, document their provenance, and prevent their destruction or abuse. Not enough is known about how the digital age impacts upon this mission, and whether resistance to computational approaches is simple risk-aversion or whether it might tell us something critical about our current conceptions of data and our present information environment.

II. Creating Narratives in the Era of Big Data

In order to navigate an information environment experiencing a 'data deluge', we seek ways to reduce noise and enhance signal, most obviously through the use of metadata. Clearly this practice involves judgements of value to determine what is worthy of the mantle of 'signal' and what is labelled 'noise'. Archival science navigates the blurred contours of this landscape, which has always been shaped by cultural and temporal perceptions and the affordances of technology. The technologies that become part of standard practice in an archive then favour the creation of certain narratives over others. If data complexity is suppressed or left unaccounted for by those technologies, it will occupy a blind spot within the archive but if its description is too bound up with its complexity, its diverse potential uses will not be discovered. Either extreme represents a dilution of the richness of knowledge creation.

In discussing big data in relation to archives, we are interested in approaches that support the potential for data to be re-used and re-analysed in conjunction with other data that may have been collected by unrelated researchers. Such research is facilitated through the use of descriptive metadata, appropriate preservation systems, informed institutional practice, and architecture for sharing across institutions to enable discovery by diffuse researchers. Mirroring wider society, academic research is currently in thrall to big data. Funding calls offer large grants to researchers who can corral the most unlikely research interests into data-rich areas, 'consigning research questions for which it is difficult to generate big data to a funding desert and a marginal position within and outside the academy' [1]. Researchers taking on this challenge must then grapple with the socially constructed nature of datasets containing knowledge complexity that must nevertheless exemplify the gold standard of a five-star (re-)usability rating, a hallmark of epistemic authority that can only be achieved by containing some of that complexity in a black box [2]. Such flattening of nuance is described as the defining characteristic of data engineering, which leads to what McPherson [3] calls a lenticular view of knowledge.

To understand what such a turn really means for archival practice, it has been argued [4], [5] that we must clarify whether big data is genuinely being adopted as a heuristic by academic, governmental and associated actors, or if the 'myth' of 'Big Data' [6] is merely a useful discourse for those whose interests are served by the promulgation of an evangelical 'dataism' [7]. This phenomenon has parallels across society. For example, Williamson [8] analyses how the Hour of Code and Year of Code initiatives saw 'a computational style of thinking' infiltrate schools in the US and UK, which he describes as a style of thinking that 'apprehends the world as a set of computable phenomena'. Williamson draws attention to a deficit of reflexivity amongst advocates of computational approaches to social problems, which obfuscates the 'worldviews, ideologies and assumptions' of the creators of systems for processing data, black-boxing the processes that delimit data use. Berry [9] draws on Fuller [10] in pointing out that the potential for new technologies to produce and reproduce inequalities in society is not simply a matter of a 'digital divide' but is significantly influenced by the commercial roots and market values of much of this *techno-solutionist* innovation.

III. How can archivists' views of knowledge complexity help to answer questions about big data?

Archivists are uniquely placed within this discourse, with everyday practices and systems for managing collections, and the confluence of traditions of working with cultural heritage holdings and adaptation to emerging technologies, all in their purview. As such, archivists are more than a vital link in the chain through which historical data are maintained and transmitted. KPLEX contends that engaging with archivists' perspectives is fundamental to understanding the drivers behind data use and non-use and that viewing the knowledge landscape from their position of archival thinking offers insight into how the computational turn is experienced in practice and how this may render new forms of research engagement with archives.

This project therefore channels that insight to tackle questions about big data and the evolution of computational archive science. A shift towards big data approaches necessarily poses questions of how the contemporary landscape is characterised and what the custodianship of cultural heritage looks like at present. Are any aspects of archivists' roles 'hidden', as Star [11] observes of much of the 'work, practice, and membership' of socio-technical networks? Might data end up hidden between the cracks of the institution, as Vanden Daelen et al. [12] documented? Do practitioners seek to maintain or arrive at a finished model or are their ideas of completeness more akin to 'equilibrium in flow'? [13].

Grounding our understanding of archival thinking in the practices, relationships and goals of archivists helps us to gauge the utility of the many lines of enquiry about its compatibility with the computational turn. For instance, an increasing marginalisation of deductive approaches has been reported in scientific fields [5], and while it would seem unlikely that humanists and social scientists would reject deductive methods in favour of purely inductive methods, the extent to which data-driven approaches are supported by archivists may be revealing. The practitioner view of the opportunities and challenges for broader use of data that big data approaches offer has been conspicuously absent from a discourse that largely represents them as passive actors, resistant to change [14],[15].

This perception assumes the work of archivists in conventional institutions has little in common with emerging practices in discrete data archives. Ribes and Jackson's investigation of the workings of the data archive [16] describes how 'the work of sustaining massive repositories reveals only a thin slice in the long chain of coordinated action that stretches back directly to a multitude of local sites and operations through which data in their "raw" form get mined, minted, and produced. What remain at repositories are the distilled products of these field sites; behind these centers lie an even more occluded set of activities that produce those data themselves'. Extant research has not fully documented the extent to which existing metadata and practices across the sector already represent a big data approach to historical and cultural sources. By applying the lens of knowledge complexity, which approaches all cultural heritage institutions as sites of complex webs of action, we seek to avoid making the value judgements that a comparative approach risks becoming embroiled in. Through careful use of terminology that probes the activities of archivists rather than pitting their values against those of the computational turn, we may uncover subtleties of action that result in data becoming hidden, for example, in gaps between colleagues with different pieces of the metadata puzzle [12] – which an interrogation of acts of deliberate resistance would not reveal.

The myth of big data hinges on the occlusion of human intervention, which is the basis of claims that big data approaches render invisible or 'remove' 'human bias' [5]. Of course, bias is central to historical inquiry, and researchers' power to recognise and expose it is key to their epistemic authority, so we might ask: if bias is hidden is a historical approach neutered? Rather than simply being a profession that is hostile to novel forms of knowledge creation (and perhaps there are some myths around preservation and conservation at work here), archivists may well have some considered reservations about the computational turn. When acting at the site of convergence of data practices as diverse as ethnography, with its concern for making researchers' positionality explicit, and big data, which tests the boundaries of linking data collected for different purposes, surely some tension is to be expected. Rather than rely on sectoral stereotypes then, our research explores ways in which knowledge complexity might impact upon archival thinking and practice.

Indeed, previous research has suggested that archivists are constantly changing and adapting their practices and systems [12], [17] and this will continue through and beyond the era of big data. Of course, there is a risk inherent in making any change to the way in which the historical record is passed on, that breaks in the chain may cause data to become hidden. Crucially, however, new practices must allow batons to be passed to future systems that might be better able to accommodate that data, as obsolescence is inevitable. This is not a new problem. In 1946, Broadfield [18] described how classification systems cannot last forever and called for declines in technology to be properly managed to preserve knowledge, arguing that: '[all] classifications in their existing forms are destined to become dust; sensitive adjustment should enable the classifier to consign them to dust himself [sic], instead of allowing the common enemy Time to do so'. Archival practice therefore never stands still, though it may change course, and an appreciation of knowledge complexity in archives can help us understand why some paths are taken while others are left unexplored.

IV. The Secret Lives of Data

In order to understand this process, one of KPLEX's tasks is to further define a model of cultural heritage holdings as data (digital and otherwise) and investigate cultural and ethical barriers to big data approaches to historical and cultural sources. In doing so we take on board and build on findings about 'uncertainty' and 'digital disorder', that led Weinberger [19] to state that 'the real problem is that any map of knowledge assumes that knowledge has a geography, that it has a top-down view, that it has a shape'. Rather than attempting to simply trace the A-to-Z of an idealised research data life cycle then, we are interested in the black boxes that characterise the processes we are investigating. In asking why data are not used we are concerned with all factors that may lead to data becoming 'hidden' from the historical record. Our use of 'hidden' is not to imply any active choices but speaks of the result: that data are not visible to researchers who might otherwise use them. Such 'hiddenness' will necessarily take many forms on a spectrum from being less conspicuously validated, for example by a missed opportunity for duplication in a specialist archive, to being more obfuscated or 'buried' in a way that diminishes researchers' chances of discovery. KPLEX therefore seeks to apply theories that might help to explain how metadata and actions become obscured. For example, Karup and Block's [20] concept of quasi-actants supplements Latour's [2] vision of black boxes as they describe how these actors erase their traces so their work is not visible even when the black box is opened – in other words their *mediation* does not become *metadata*. Engaging with archivists' perspectives is vital for grasping the nature and effects of this hidden work.

From a deeper understanding of the daily manifestations of archival thinking about data, we can forensically inspect the definitions and goals of institutions, relationships and practices that actors subscribe to. A less than harmonious agreement about these might help to explain the likelihood of any gaps that undermine the passage of data from creation to re-use. Crucially, in turning the focus to human factors as more than a blockage in the pipeline, we can get to the root of any genuine concerns archivists might have. It may be that big data approaches are seen as a fundamental shift or re-purposing of the archive, with practitioners feeling less like they are being nudged at the micro level and more like they are being 'enrolled' [2] by discourses of data science or commercialisation at the macro level. Latour and Callon described how translating the terms of a problem from the language of one discipline to another achieved *intéressement* when the translation is maintained and reinforced in order to complete the transfer of power from one set of actors to another [11]. Concepts such as Lave and Wenger's [21] legitimate membership of communities of practice might be applied to describe experiences of translation. With some observers suggesting that academia is in the grip of an intellectual land-grab [22], do knowledge practitioners fear becoming a McArchive? In narrowing our focus down to practitioners' experiences of complexity as custodians of public knowledge, we hope to give meaning to grand narratives on themes like the privatisation of knowledge.

Simply observing that many data scientists enjoy finding patterns in numbers and many historians are motivated by a passion for telling the stories of people who have suffered in being reduced to numbers does little to progress debate or practice in either field. It is nevertheless instructive to contrast a commitment to learning from extraordinary past events as a typical feature of an archive's mission statement to McPherson's [3] analysis of the lenticular view of UNIX-style structures for coding, in which complexity is managed and controlled through the 'principles of information hiding' and the creation of discrete modules devoid of relation and context. McPherson highlights the benefits of such a modular approach for coding, while warning that it also represents 'a worldview in which a troublesome part might be discarded without disrupting the whole' [3]. This approach threatens to engineer apophenia – 'seeing patterns where none actually exist, simply because enormous quantities of data can offer connections that radiate in all directions' [6] – creating an information environment in which (potentially erroneous) macro-level patterns govern our view of knowledge creation. As well as reservations about offering up data to abuse or putting a balanced understanding of the past at risk, practitioners may have fundamental ethical fears about data linking that stem from their professional knowledge of the potential use of their collections. If practitioners are not convinced that privacy and research integrity can be maintained when datasets are linked or have concerns about ownership, control or access, now is our opportunity to rigorously interrogate these issues so that research moves forward consciously, purposefully and without opening Pandora's box.

V. Emerging Themes

Previous research [15] has suggested resistance to greater data sharing among cultural heritage practitioners, which has been taken as a proxy for a rejection of big data as both myth and heuristic, but nuance is important here. Where change is resisted, is it the novelty itself or a perceived 'translation'? Of course, commonalities exist between 'conventional' archival practice and that of the data archive or sharing initiative, so which concepts and values are accepted as shared and which are disputed? How is this ownership and epistemological agency experienced and acted out and where exactly does the no man's land of unresolved differences lie?

Our early findings suggest that some kind of buffer between archival thinking and computational thinking may be a helpful part of this process. One archivist explained that, after overcoming her initial reservations about adapting to new technologies and seeing the need to "speak IT", she adopted the practice of using an in-house IT colleague to "transfer the message" to the software company. When it comes to infrastructure projects, one of our participants described their utility as allowing her to "connect to different worlds" but cautioned that such platforms were "tools, not solutions". Again, this approach challenges a simplistic understanding of archivists' relationships with new technologies and highlights the need to understand their agency in a process that must accommodate the complexity of materials and institutions. By looking at (and for) knowledge complexity, we are taking a pragmatic approach to cut through some of the more sensational readings of the current data trends. Emerging findings from KPLEX suggest that archivists do perceive, and are adapting to, a shift in researchers' methods. One participant felt that the hierarchy of collections was no longer an important source of context as it is not relevant to the search methods that now predominate researchers' practice. For her, acknowledging this phenomenon meant putting greater onus on collection descriptions for providing context. Another archivist observed that a 'hidden' aspect of his role was the importance of following a model for holdings descriptions. His experience suggested that consistent completeness of metadata was a long under-appreciated necessity for the representation of complexity. Although this participant was an advocate of archivists' role in informing researchers of context and wary of the "quick wins" of search methods, it was an understanding of the requirements for working across complex collections, rather than a new need to accommodate computational approaches that had led him to 'respect' standardisation. KPLEX is therefore uncovering some of the ways in which elements of knowledge complexity and big data approaches coalesce, and we are beginning to unpick what really drives continuity and change in archival practice.

So far, we have found no evidence of the 'digital drama' Kitchin [5] describes in terms of a 'new empiricism' with the logic that 'through the application of agnostic data analytics the data can speak for themselves free of human bias or framing, and any patterns and relationships within Big Data are inherently meaningful and truthful'. Clearly, this is not a research model that we would expect archivists to gear their services towards. Indeed, fear of apophenia cannot be said to have grown out of, let alone be confined to, the humanities. Rather, the humanities are positioned at the extreme of most resistant to big data because their concern for the human – that is, the dangers of espousing relationships with no human framing – is demonstrably acute. What we have found so far, however, are examples of archivists looking for opportunities to embrace those elements of the computational turn they see as beneficial and weave them into their practice without distorting the fabric of the archive. KPLEX therefore aims to make sense of interpretations of big data in archives, helping to chart the barriers and bridges between archival thinking and computational thinking by looking at and for knowledge complexity. By understanding the complexities of archivists' habitual practice and thinking, we can do more than stay afloat in the data deluge. Just as Manovich [23] laments the information trend of telling only the middle of the story so that '[r]ailway trains only begin to exist when they are derailed', with what came before and after the defining event consigned to the category of 'noise', so might we obtain more meaning from the 'deluge' if we are able to appreciate the tributaries that feed it and the dams it is most likely to burst.

Acknowledgement

KPLEX is funded by the European Commission's Horizon 2020 Programme, Contract Number 732340.

References

[1] R. Kitchin, The Big Data Revolution. London: SAGE, 2014, p. 28.

[2] B. Latour, B, Pandora's Hope: essays on the reality of science studies. Cambridge: Harvard University Press, 1999.

[3] T. McPherson, *Why Are the Digital Humanities So White? or Thinking the Histories of Race and Computation*, Debates in the Digital Humanities, 2012. [Online]. Available: http://dhdebates.gc.cuny.edu/debates/text/29 [Accessed: Oct. 1, 2017].

[4] G. Bolin and J.A. Schwarz, "Heuristics of the algorithm: Big Data, user interpretation and institutional translation," Big Data & Society, vol. 4, July–Dec. 2015, pp. 1–12.

[5] R. Kitchin, "Big Data, new epistemologies and paradigm shifts," Big Data & Society, vol. 1, April–June 2014, pp. 1–12.

[6] d. boyd and K. Crawford, "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon," Information, Communication, & Society, vol. 15, May 2012, pp. 662-679.

[7] J. van Dijck, "Datafication, Dataism and Dataveillance: Big Data between scientific paradigm and ideology," Surveillance & Society, vol. 12, May 2014, pp. 197–208.

[8] B. Williamson, "Political Computational Thinking: policy networks, digital governance and 'learning to code'," Critical Policy Studies, vol. 10, June 2015, pp. 39-58.

[9] D.M. Berry, "The Computational Turn: thinking about the digital humanities," Culture Machine, vol. 12, 2011, pp.1-22.

[10] S. Fuller, "Humanity: The Always Already – or Never to be – Object of the Social Sciences?," in *The Social Sciences and Democracy*, Ed. W. Bouwel, London: Palgrave, 2010.

[11] S.L. Star, "Power, Technology and the Phenomenology of Conventions: on being allergic to onions," in: *Technoscience: The*

Politics of Interventions, Ed. K. Asdal, B. Brenna, and I. Moser, Michigan: Unipub, 2007, p. 95.

[12] V. Vanden Daelen, J. Edmond, P. Links, M. Priddy, L. Reijnhoudt, et al. "Sustainable Digital Publishing of Archival Catalogues of Twentieth-Century History Archives," Open History: Sustainable digital publishing of archival catalogues of twentieth-century history archives. Brussels, Belgium, Dec. 2015.

[13] L. von Bertalanffy, Vom Molekül zur Organismenwelt: Grundfragen der modernen Biologie. Athenaion, 1949.

[14] J.J. Duderstadt, D.E. Atkins, J. Seely Brown, M.A. Fox, R.E. Gomory, N. Hasselmo, P. Horn, et al., Preparing for the Revolution: Information Technology and the Future of the Research University, Washington, DC: National Academies Press, 2002.

[15] P.N. Edwards, S.J. Jackson, M.K. Chalmers, G.C. Bowker, C.L. Borgman, D. Ribes, M. Burton and S. Calvert, S. Knowledge Infrastructures: Intellectual Frameworks and Research Challenges, Ann Arbor: University of Michigan School of Information, 2013.

[16] D. Ribes and S.J. Jackson, "Data Bite Man: The Work of Sustaining a Long-Term Study," in: *Raw Data" is an Oxymoron*, Ed. L. Gitelman, Cambridge: MIT Press, 2013, pp. 147-166.

[17] C.L. Borgman, Big Data, Little Data, No Data: Scholarship in the Networked World, Cambridge: MIT Press, 2015.

[18] A. Broadfield, The Philosophy of Classification, London: Grafton, 1946, pp. 65-66.

[19] D. Weinberger, Everything Is Miscellaneous: The Power of the New Digital Disorder, New York: Holt. 2007, p. 63.

[20] T. Karup and A. Block, "Unfolding the Social: Quasi-actants, Virtual Theory, and the New Empiricism of Bruno Latour," The Sociological Review, vol. 59, March 2011, pp. 42-63.

[21] J. Lave, and E. Wenger, Situated Learning: Legitimate Peripheral Participation, Cambridge: Cambridge University Press, 1991.

[22] C. Hess, and E. Ostrom, "Ideas, Artifacts, and Facilities: information as a common-pool resource," School of Law, Duke University, vol. 66 Winter/Spring 2003.

[23] L. Manovich, The Exceptional and the Everyday: 144 Hours in Kiev, IEEE International Conference on Big Data, Washington DC Oct. 2014.