DHN 2018 Abstract: When Open becomes Closed: Findings of the Knowledge Complexity (KPLEX) Project.

Authors: ***Jennifer Edmond, Nicola Horsley, Elisabeth Huber, Georgina Nugent-Folan, Rihards Kalnins, Jörg Lehmann, Mike Priddy, Thomas Stodulka, and Andrejs Vasiljevs.***

Submitted and presented by:
Georgina Nugent-Folan (nugentfg@tcd.ie)

Relevant themes: Open Science, Cultural Heritage & The Future.

The future of cultural heritage seems to be all about "data." A Google search on the term "data" returns over 5.5 billion hits, but the fact that the term is so well embedded in contemporary discourse does not necessarily mean that there is a consensus as to what it is or should be. The lack of consensus regarding what data are on a small scale acquires greater significance and gravity when we consider that one of the major terminological forces driving ICT development today is that of "big data." So too do terms such as "data cleaning," "signal" and "noise." While the phrase may sound inclusive and integrative, "big data" approaches are highly selective, excluding any input that cannot be effectively structured, represented, or, indeed, digitised. The future of DH, of any approaches to understanding complex phenomena or sources such as are held in cultural heritage institutions, indeed the future of our increasingly datafied society, depends on how we address the significant epistemological fissures in our data discourse. This is not to say that digital data analysis approaches cannot also nurture epistemic multiplicity, but that there are observable biases to be found in some aspects of big data research. For example, how can researchers claim that "when we speak about data, we make no assumptions about veracity"[1] while one of the requisites of "big data" is "veracity"?[2] On the other hand, how can we expect humanities researchers to share their data on open platforms such as the European Open Science Cloud (EOSC) when we, as a community, resist the homogenisation implied and required by the very term "data," and share our ownership of it with both the institutions that preserve it and the individuals that created it? How can we strengthen European identities and transnational understanding through the use of ICT systems when these very systems incorporate and obscure historical biases between languages, regions, and power elites? In short, are we facing a future when the mirage of technical "openness" actually closes off our access to the perspectives, insight, and information we need as scholars and as citizens? How might this dystopic vision be avoided?

These are the questions and issues under investigation by the European Horizon 2020 funded Knowledge Complexity (KPLEX) project,[3] and we are doing so by applying strategies developed by humanities researchers to deal with complex, messy, cultural data; the very kind of data that resists datafication and poses the biggest challenges to knowledge creation in large data corpora environments. Arising out of the findings of the KPLEX project (the conclusion of which is fortuitously coterminous with DHN2018), this paper presents the

[1] Rosenberg, "Data before the Fact," in ibid., 37.
[2] See the cfp of BDIOT 2017 (http://www.bdiot.org/cfp.html) and IEEE 2017 (http://cci.drexel.edu/bigdata/bigdata2017/) respectively.
[3] The KPLEX project has been funded by the European Commission's Horizon 2020 Programme, Contract Number 732340.

synthesised findings of an integrated set of research questions and challenges addressed by a diverse team led by Trinity College Dublin (Ireland) and encompassing researchers in Freie Universität Berlin (Germany), DANS-KNAW (The Hague) and TILDE (Latvia). As this paper will make clear, we have adopted a comparative, multidisciplinary, and multi-sectoral approach to addressing the issue of bias in big data; focussing on the following four key challenges to the knowledge creation capacity of big data approaches:

1. Redefining what data is and the terms we use to speak of it (TCD);
2. The manner in which data that are not digitised or shared become "hidden" from aggregation systems (DANS-KNAW);
3. The fact that data is human created, and lacks the objectivity often ascribed to the term (FUB);
4. The subtle ways in which data that are complex almost always become simplified before they can be aggregated (TILDE).

This paper presents a synthesised version of these integrated research questions, combining qualitative and quantitative approaches to discuss the overall findings and recommendations of the project. What follows gives a flavour of the key issues addressed by the four project teams, and the related project findings that will be presented throughout this paper.

1. Redefining what data is and the terms we use to speak of it. Many definitions of data, even thoughtful scholarly ones, associate the term with a factual or objective stance, as if data were a naturally occurring phenomenon.[4] But data is not fact, nor is it objective, nor can it be honestly aligned with terms such as “signal” or “stimulus,” or the visceral but misleading “raw data.” To become data, phenomena must be captured in some form or agent, signal must be separated from noise and like must be organised against like. In short, transformations occur. These organisational processes are either human determined or human led, and therefore cannot be seen as wholly objective; irrespective of how effective a (human built) algorithm may be. The core concern of this keystone facet of the project was to expand extant understanding of the heterogeneity of definitions of data, and the implications of this state of understanding. By establishing a clear taxonomy of existing theories and definitions of data—identifying the key terms (and how they are used differently), key points of bifurcation, and key priorities under each conceptualisation of data—we provide a foundation which serves to underpin a more applied tautology of humanistic versus technical applications of the term. Our taxonomy of data definitions is backed up by the findings of a data mining exercise wherein we examined usage of the terms “data” and “big data” in the proceedings of major international big data journals from their inception to the present day. The overarching insights obtained by the taxonomy of data definitions and the data mining exercise are counterbalanced by the findings of thirteen in depth interviews with computer scientists, conducted with a view to obtaining a more detailed picture of the various understandings of the term “data” that underlie computer science research and development.

2. Dealing with “hidden” data. According to the 2013 ENUMERATE Core 2 survey, only 17% of the analogue collections of European heritage institutions had at that time been

[4] See for example Rowley, Jennifer. (2007) "The Wisdom Hierarchy: Representations of the DIKW Hierarchy." Journal of Information Science 33, no. 2: 163–80.

digitised.[5] This number actually represents a decrease over the findings of their 2012 survey (almost 20%). The survey also reached only a limited number of respondents: 1400 institutions over 29 countries, which surely captures the major national institutions but not local or specialised ones. Although the ENUMERATE Core 2 report does not break down these results by country, one has to imagine that there would be large gaps in the availability of data from some countries over others. Because so much of this data has not been digitised, it remains "hidden" from potential users. This may have always been the case, as there have always been inaccessible collections, but in a digital world the stakes and the perceptions are changing. The fact that so much other material is available on-line, and that an increasing proportion of the most well-used and well-financed cultural collections are available digitally as well, means that the reasonable assumption of the non-expert user of these collections is that what cannot be found does not exist (whereas in the analogue age, collections would be physically contextualised with their complements, leaving the more likely assumption to be that more information existed, but could not be accessed). The threat that our narratives of histories and national identities might thin out to become based on only the most visible sources, places and narratives is high. Through studying the often-neglected perspective of the archivist, this facet of the project elucidated the manner in which data that are not digitised or shared become "hidden" from aggregation systems and how cultural heritage practitioners are working to overcome the practical challenges that underlie ambitions of expanding access to public knowledge.

3. Knowledge organisation and the epistemics of emotions data. The nature of humanities data is such that even within the digital humanities, where research processes are better optimised toward the sharing of digital data, sharing of "raw data" remains the exception rather than the norm. The "instrumentation" of the humanities researcher consists of a dense web of primary, secondary and methodological or theoretical inputs, which the researcher traverses and recombines to create knowledge. This synthetic approach makes the nature of the data, even at its "raw" stage, quite hybrid, and already marked by the curatorial impulse that is preparing it to contribute to insight. This aspect may be more pronounced in the humanities than in other fields, but the subjective element is present in any human triggered process leading to the production or gathering of data. Emotions and affects serve as an effective cross-disciplinary topic, because few phenomena are as tricky in terms of datafication and measurement. This facet of the project elucidates that there is no shared vocabulary on either affect or emotion, nor data itself, with regards to differing "epistemic cultures."[6] By means of a series of in depth interviews and an online survey we investigated the researchers' view on affects' and emotions' resistance to datafication, and thus what escapes the access of positivistic approaches. Difficulties arise especially in the translation of scientific claims to objectivity[7] into research methodologies. Theoretical and methodological biases are inevitable when doing research on emotions and affects, and datafication therefore leads to a reduction and marginalization of the complexity of the research object "emotions." The insights gained will make visible many of the barriers to the inclusion of all aspects of science under current Open Science trajectories, and reveal further central elements of social and cultural knowledge that are unable to be

[5] http://www.enumerate.eu/en/statistics

[6] Karin Knorr-Cetina, Epistemic cultures: How the sciences make knowledge, Cambridge, Mass. u.a.: Harvard Univ. Press 1999.

[7] Lorraine Daston, Peter Galison, Objectivity, New York: Zone Books 2007.

accommodated under current conceptualisations of “data” and the systems designed to use them.

4. Cultural data and representations of system limitations. Cultural signals are ambiguous, polysemic, often conflicting, and contradictory. The process of “data-fication” robs culture of their polysemy, or at least reduces it to elements that have been, or can be, classified, divided, and filed into taxonomies and ontologies. One of the greatest challenges for so-called “big data” is the analysis and processing of multilingual content. This challenge is particularly acute for unstructured texts, which make up a large portion of the “big data” landscape. The language technology (LT) industry serves as an ideal test case for examining issues surrounding data, its availability, and the impact of data on technology, infrastructure, and employment. LT solutions are developed with language data as input material, therefore data issues—e.g., errors, noise, and inconsistencies in coverage—have a crucial impact on service quality. Input data issues become more acute when neural networks are used in development. AI-based solutions like Neural Machine Translation (NMT) are more sensitive to input data mistakes, often treating them as linguistic phenomena. Mistakes are exacerbated by data scarcity, particularly for smaller languages and overlooked domains. To understand the impact of data inconsistencies on LT, in this facet of the project we analyzed data availability for EU languages, including large-scale corpora, multilingual open data, and resources available for the CEF eTranslation platform, an analysis that has led to several hypotheses, subsequently validated by a survey of LT experts and EU language community representatives. In addition, we recount our efforts to analyze the effects of data inconsistencies on Neural MT. By exploring LT as a test case, we intended to show how data inequality will potentially become a major theme in “big data.” Furthermore, once it has been examined in the context of LT and “big data,” the overarching social and political consequences of data inequality will become apparent, helping to inform possibilities for policy decisions on the part of EU institutions.