Edmond, Jennifer, Nicola Horsley, Jörg Lehmann, and Mike Priddy. "Viewing big data through the lens of culture." *The Trouble With Big Data: How Datafication Displaces Cultural Practices*. London,: Bloomsbury Academic, 2022. 1–19. *Bloomsbury Collections*. Web. 26 Apr. 2023. <<http://dx.doi.org/10.5040/9781350239654.ch-1>>.

Downloaded from Bloomsbury Collections, www.bloomsburycollections.com, 26 April 2023, 13:34 UTC.

Copyright © Jennifer Edmond, Nicola Horsley, Jörg Lehmann and Mike Priddy 2022. Released under a CC BY-NC-ND licence (<https://creativecommons.org/licenses/by-nc-nd/3.0/>). You may share this work for non-commercial purposes only, provided you give attribution to the copyright holder and the publisher, and provide a link to the Creative Commons licence.

Viewing big data through the lens of culture

The development of this book began, as so many good stories do, with a group of people locked in a conflict. What made this conflict all the more intense was the fact that everyone in the room agreed what needed to be done, and yet none of them was satisfied with the outcome of previous attempts to meet the agreed requirement. They needed to share data.

The people in question were an interdisciplinary group of researchers, including historians, computer scientists and collections experts, trying to build an innovative platform to support historical research. This aspect of the project had largely stalled, due to what we ultimately recognized as a malfunction in communication. In essence, each party was working from a completely different definition of the word data, and a different understanding of what was possible and available as data to drive and populate the system.

In fairness, this problem actually began well outside of the group that had gathered, as the agency that had funded the project had done so already under something of a misconception about historical data, namely that historical resources were largely already available in digital forms, and merely needed to be somehow aggregated to facilitate their use. The historians in the group were of course deeply suspicious of this: they were keenly aware that only a fragment of the resources they needed were digitized, and greatly unevenly at that. They also weren't really confident about the word data, as it wouldn't be the word they would normally apply to their sources. But they took the impression from their technical colleagues that this word was somehow coterminous with 'inputs' and therefore began to offer small, rich, heterogeneous or indeed analog collections as examples. What the technical development team needed and wanted, however, was something more akin to 'big data': relatively homogeneous, relatively well-prepared, machine-readable full text able to support investigation of complex research questions. The collections experts from cultural heritage institutions, however, knew that realistically what was available and processable was not the research data itself, but rather the finding aids and collections metadata.

In retrospect, it is easy to see how these three different communities and perspectives would bring completely different conceptualizations of the word data to this collaboration, but at the time, this source of disconnection proved hugely challenging to tease out and manage. This left some members of the project team to wonder: if this smart, but diverse, group could end up so stymied by the gaps between their differing understandings of what the word data might encompass, was it possible that this same issue played out in other projects, sectors and arenas of life as well?

The KPLEX project

A second project was devised to test this hypothesis, and its findings form the core of the evidence presented in this book. This project, called ‘Knowledge Complexity’ (or KPLEX for short) was constructed to seek out these kinds of gaps and biases in research based on big data, harnessing in particular a multidisciplinary, but culturally centred, approach to this challenge. Specifically, the KPLEX project was moulded around three potential sources of miscommunication within teams undertaking data-related research, gaps that could also be found disrupting progress for research projects, as well as for the individual and industrial users some of this work was ultimately destined to serve. This phenomenon, whether unnoticed, convenient or intentional, resulted in what one of the authors of this book referred to as a game being played of mutual misunderstanding, according to which all sides may feel they have won, but only because they are using different rule books.

To approach this subtle and deeply engrained challenge, we defined three strands to our approach: In the first of these, we determined to look at issues of language and discourse around data, the heterogeneity of definitions of data, and the implications of this state of multiple, faceted common understandings. In particular, we also looked at the perceived tensions between data and narrative, as the building blocks of two different sensemaking techniques (one on the ascendant, one seemingly as old as our species), and the contrasts in these processes in their management of the iterative and nonlinear aspects of knowledge creation. We hoped that in this way we would be able to capture not just what the computational approach to knowledge creation might flatten, but also the concomitant processes developed not necessarily by humanists, but by humans, to preserve and make meaning from these noisy signals.

Within this strand, we particularly leveraged methods and perspectives from literary studies and historical research. For example, provenance is a key concept for historians and collections management professionals: indeed, a piece of evidence loses its authority utterly if its provenance is not clear. But in big data systems, provenance data is more likely to be looked upon as noise than signal. This is an understandable response to the problem of complexity and volume, even in the realm of cultural data: indeed, as a recent newspaper article puts it, ‘the digitization of cultural heritage is booming, but it happens to be messy and chaotic. Scientists therefore cannot find or search through numerous sources.’¹ Cultural signals are ambiguous, polysemic, often conflicting and contradictory. This is true even in the ‘low context’ cultures (see the discussion of this characteristic in Chapter 8), where a greater cultural permeability is facilitated by explicitness in the communication and day-to-day deployment of cultural norms and practices. This is inscribed most visibly in language, but also in personal interactions, in religious practices, and in artistic production. In order to transform culture into data, its elements – as all phenomena that are being rendered as data – have to be classified, divided and filed into taxonomies and ontologies. This process of ‘data-fication’ thereby robs them of their polysemy, or at least reduces it.

The question remains open, however, of what the scientists referenced above would do with data they might find in a simplified system, once they found it, however, as the simplification that comes with greater aggregation and interrogability is also an impoverishment in terms of depth and complexity. Digging into these differences, the manner in which provenance is preserved in some systems, the nature of what provenance means for research: all of this is yet to be described and expressed in a manner that can assist developers to better enable similar transparency and flexibility in data-driven research to what historians and related disciplinary communities have developed over time.

It is difficult to speak of historical research only through the lens of the historian, however, as the interdependency between cultural research and the cultural heritage institution is a long-standing and productive one. For this reason, the KPLEX project also looked at the production of catalogue metadata for objects: as holdovers from a pre-digital age of physical catalogues, as the most common data to be found in digital systems of cultural data, as structured data of a sort that is easy to aggregate, as a draw on cultural institution resources that must create it in their digital systems, and as marks of human interpretation and occasional error. In the words of Johanna Drucker: ‘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 organize and present that knowledge in digital form.²

You cannot look at these established practices in the context of big data, however, without also looking into how emerging computational and information retrieval approaches, such as ultra-large systems and deep learning, but also things as mundane as keyword search, may be displacing the production of such metadata, removing the human investment and replacing it with a proxy that may or may not serve quite the same function. It was this trend towards knowledge loss that formed our second strand of investigation. According to the 2017 ENUMERATE Core 4 survey, only 22 per cent of the analog collections of European heritage institutions had at that time been digitized.³ The survey reached only a limited number of respondents: less than 1,000 institutions over twenty-eight countries (including respondents from countries outside Europe, which were excluded from the analysis), which surely capture the major national institutions but not local or specialized ones.⁴ Although the ENUMERATE report does not break down these results by country, one has to imagine that there would be large differences in the availability of data from some countries over others. Because so much of this data has not been digitized, 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 online, and that an increasing proportion of the most well-used and well-financed cultural collections are 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 analog age, collections would be physically contextualized 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.

Data is not only hidden from the aggregated, on-line view because it has not been digitized, however. Increasingly, users are becoming frustrated with digital silos. The current paradigm is not that one visits a number of news or information sites, but that one channels one's content through an intermediary, such as Facebook or Twitter. The increase in the use of syndicated feeds such as RSS (Really Simple Syndication), or data access modalities that go around the human-readable interfaces such as APIs (Application Programming Interfaces), among many other technologies (including personalization and adaptation algorithms), evidences this preference. Cultural heritage institutions (CHIs)

have adapted to this paradigm shift by establishing their own curated spaces within these channels, but in spite of this ‘pushing out’ response, the vast majority of their data cannot yet be ‘pulled in’ by developers wanting to feature cultural content. The biggest exception to this rule in Europe is Europeana, which has a very popular API and makes the metadata it delivers available under an open CC-0 reuse licence.⁵ Most national, regional or local institutions hesitate to do the same, however, in part because of technical barriers, but also to a great extent because they do not trust the intermediaries and re-use paradigms that are emerging. They also may themselves not be in possession of full knowledge regarding the nature and import of what they hold, as the traditional symbiosis between the users of collections (i.e. the historians and other researchers) was based upon a division of knowledge, which has never been systematically integrated back into the CHIs. These institutions have developed over centuries to protect the provenance of items in their care, and to prevent their destruction or abuse. Not enough is known about how the digital age impacts upon this mission, and whether the hesitation to release data into shared platforms is merely risk-aversion, or whether this can tell us something critical about our current conceptions of data, and our current data-sharing environment.

This investigation was underpinned by the concept of the sharing of cultural data as a part of the institutions’ ‘public task’. This is a key concept, as it defines not just the generic mission of the institution, but also their responsibilities under the Public Service Information Directive to make the information they hold open.⁶ The 2012 discussion of the position of the CHIs under this directive began to define some of the parameters for what KPLEX would investigate: the fact that CHIs often don’t ‘own’ their own content in a clear way, the heterogeneity of descriptions across types of institutions and collections, the conflict between business models privileging digital access and the wider concepts of public good, among others.

Finally, the KPLEX project’s third area of focus was to take a comparative approach to a range of research processes, humanities and cultural studies among them, in order to understand the range of strategies professional researchers were deploying in the integration of data into their knowledge creation processes. In February 2016, the European Commission released a list of actions under the heading of a ‘Draft Open Science Agenda’.⁷ Many of the actions listed there, such as increasing the reliability, efficiency and responsiveness of scientific enquiry, would be wholeheartedly supported by humanists and social scientists as well as by physical or natural scientists. But the nature of humanities data is such that even within the digital humanities, where research processes are better

optimized towards the sharing of digital data, sharing of 'raw data' remains the exception rather than the norm.

There are a number of reasons for this. First of all, in many cases, ownership of the underlying input data used by humanists is unclear, and therefore the question of what can be shared or reused is one which the individual researcher cannot independently answer (this issue will be dealt with in the context of the topic described above). There are deeper issues, however, based in the nature of the epistemic processes of the humanities that act as further barriers to reuse of humanities data. Very little research exists in this topic to date, although barriers to reuse of digital humanities projects do provide an interesting baseline for starting an investigation. For example, the LAIRAH project pointed towards a number of key issues leading to a lack of reuse of digital data prepared by research projects.⁸ In particular, the lack of an early conceptualization of who the future user of the data might be and how they might use it was a key deterrent to future use. While this lack may be seen as a weakness from a reuse standpoint, it is likely that the organization of data or the curation of resources chosen in such projects was driven by the research questions in the mind of the original researcher, and that this organizational model was key to their epistemic process. As the results of research into humanistic research processes⁹ have demonstrated, 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, somewhat 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 individual element is present in any human-triggered process leading to the production or gathering of data. Another element of this is the emotional. Emotions are motivators for action and interaction that relate to social, cultural, economic and physiological needs and wants. Emotions are crucial factors in relating or disconnecting people from each other. They help researchers to experientially assess their environments, but this aspect of the research process is considered taboo, as noise that obscures the true 'factual signal', and as less 'scientific' (seen in terms of strictly Western colonialist paradigms of knowledge creation¹⁰) than other possible contributors to scientific observation and analysis. What we need is more conceptual clarity when we assess global knowledge and translate it into federated, shared digital datasets. We used this frame of reference, that of emotion researchers, to explore the data creation processes

of the humanities and related research fields to understand how they combine pools of information and other forms of intellectual processing to create data that resists datafication and 'like-with-like' federation with similar results. In particular, we were interested in locating 'failure points' in data, that is, points at which data becomes unusable, what characterizes these points and what a researcher does about it.

In sum, while we did not set out to position KPLEX as a digital humanities project, ultimately we seem to have done so, harnessing the emerging perspectives of critical digital humanities approaches¹¹ to turn eyes attuned to both cultural and humanistic forces at play as well as the impact of engineered knowledge technologies towards objects of study outside of the humanities themselves. Though we drew from the work and methods of social science, information science and computer science, our core desire from the outset was to look at the cultures, languages, values and beliefs at play below the surface of big data research, and this also comprises what we see as the primary contribution of the project.

The KPLEX interviews

The team that investigated these issues within the KPLEX project harnessed a number of methods, including an extensive engagement with secondary research across a number of fields, surveys, data-mining exercises, participation in relevant meetings and presentations and so on. Of particular importance, however, was the production of a corpus of interviews, thirty-eight in total, each of approximately one hour in length. The subjects of these interviews were roughly evenly split between participants representing three perspectives: computer scientists working with cultural data, interdisciplinary researchers, all focused on human emotion but with different home disciplines and levels of data intensity in their approaches, and cultural heritage professionals charged with preserving and making cultural source data accessible. In each cohort, the selected participants' work had a cultural element but was conceived of in terms of other disciplines (computer science, neuroscience, computational linguistics, etc.), or, in the case of the cultural heritage professionals, crossed boundaries between a number of disciplines. Although every group was interviewed with the intention of exploring their attitudes and practices towards and within big data research environments, the interview protocols used were tailored in each case to the professional discourse and specific concerns of that cohort, as well

as to the facet of the project's research questions their input was most likely to shed light on.

Anonymized transcripts of these interviews are openly available as research data and they will be cited extensively in the course of this book.¹² It is an understatement to say that they provide a valuable and unique perspective on some of the darker sides of data-driven research, its propensity to incorporate biases of the teams and processes that assemble and use it, the risks it poses to individuals and cultural codes and the manner in which it tends to obscure other forms of sensemaking, when we in fact need more, not fewer, ways of approaching complex problems.

These results were collected, analysed and filtered through the perspectives of the full KPLEX team, which included researchers steeped in the practices of literary studies, history, anthropology, sociology, information science, software development and the digital humanities. As such, this team was able to bring a fresh perspective to existing work emerging from science and technology studies, looking at not only how big data research and development practices shape our economy and society, but how it becomes entangled with fundamental aspects of our culture: how we talk, what we accept and rebel against, whom we believe, how we come to understand the world and our place in it, who we think we are (and who we think others might be), the evidence we base these identities and cultural memories on, how we build hierarchies and social capital and so forth. We were able to apply this view from another perspective, and indeed to use our position and those of many of our interview subjects, as professionally trained seekers of knowledge, to explore these specialized processes, and the manner in which biases, quirks and flaws in them seemed to trickle down into wider social and cultural challenges.

Knowledge complexity 'in the wild'

The generation, gathering, analysis and interpretation of data have been central to knowledge creation and claims to truth since the first academic disciplines began to emerge. Research that makes use of data that can be seen to 'speak for themselves' has been known to leave a lasting mark on our culture; such is their capacity for capturing the popular imagination. Just as the Covid-19 pandemic has focused minds on a singular goal, eclipsing other priorities, widespread outbreaks of cholera in the 1800s led to concerted efforts in tracking the spread of disease, including John Snow's contamination analysis and pioneering

maps – an early example of accessible data visualization that crossed into the social sciences with Charles Booth's colour coding of neighbourhoods by social class. In spite of the underlying data-centrism of Booth's iconic visualization, however, it is important to note that this was as much an investigation of culture as of social or indeed epidemiological evidence about communities, habits and trust as well as about bacteria and poverty.

Such iconic representations of data have come to stand for the progress made by our most celebrated innovators across the academic spectrum. Recent digitization initiatives have sought to expand the potential of facilitating the uptake of data-driven methods for areas of study previously seen as necessarily resistant. Arts, humanities and social science methods in particular were seen as intrinsically concerned with issues the digital did not easily encompass, such as close examination of the provenance and unique qualities of sources, considering them in their original contexts, looking for rich exemplars rather than large statistical patterns, slow processes and the fundamentally analog. This digital turn is part of a general trend whereby the research community and society at large have come to view data themselves as offering the answers – and even the questions – that bring us closer to understanding ourselves than the conventional methods of the humanities and social sciences. The reputation of the scientific method as a gold standard and perceptions of the essential 'purity' of data themselves seem invulnerable to even high-profile scandals. Whether duped by the publication of a counterfeit data-linking exercise in *The Lancet*¹³ or the havoc wreaked by a 'mutant algorithm' deciding the grades by which young people's admission to university would be decided,¹⁴ the tendency is to decry a wasted opportunity to properly harness the revealing powers of the data as always more objective than a fallible human counterpart decision-making process (in spite of the similar levels of bias and inappropriate modelling possible in each paradigm).

And so data have come to be seen as 'reusable goods',¹⁵ their value no longer subject to interpretation by specialists in a particular discipline but dependent on their malleability to serve new, unpredicted purposes as they are aggregated with other data plucked from disparate contexts. This divergence from knowledge norms exemplified by Bernstein's *bounded knowledge code*¹⁶ disrupts the process of knowledge creation and discovery, recalibrating its fundamental dynamics away from archival classification values that have been optimized for the needs of professional historical researchers, through a shift in practices located in institutional, governmental and commercial spheres. The standardization of data that began with trade between food markets accelerated with epidemics and

facilitated globalization. It was put to use in building global infrastructures for sharing digitized data from most scientific disciplines from the 1970s, expanding through the ubiquity of personal computers and faster processing systems to our current demand for instant query fulfilment. Big data's hegemonic position has been cemented by the favour it has found in funding priorities and the policies and practices they have reinforced and engendered.

Sceptics of this hegemony, such as Donna Haraway, have instead advocated 'situated knowledges'. 'Situated knowledges are about communities, not about isolated individuals.' Haraway commented that '(i)mmortality and omnipotence are not our goals', nor is the splitting of subject and object.

But we could use some enforceable, reliable accounts of things not reducible to power moves and agonistic, high-status games of rhetoric or to scientific, positivist arrogance. This point applies whether we are talking about genes, social classes, elementary particles, genders, races, or texts; the point applies to the exact, natural, social, and human sciences, despite the slippery ambiguities of the words 'objectivity' and 'science' as we slide around the discursive terrain.¹⁷

And yet, the manner in which data are collected, used, reused, conceptualized and enshrined in language seems almost to demand and insist upon occupying a status above situatedness, a contextless state in line with the omniscience and omnipotence to which they seem to be granted a claim. (W. Edwards Deming was famously credited with coining the phrase: 'In God we trust. All others, bring data.') But we are our data, it is created not given, and therefore it is, for all the veneer of objectivity its champions want to place upon it, as fallible as we are and bound to human ideation, hierarchies, values and imaginaries – in other words to our culture. This obstacle, and the frustration that resulted in trying to work within its shadow as researchers that had to regularly transverse the boundaries between data-driven exploration and the qualitative traditions of cultural studies to pursue their objects of study, was the inspiration for this book.

As we shall see in Chapter 2's consideration of the linguistic images used to represent the idea of data, there is a disconnect between the enthusiasm for data as a panacea and the struggle of the popular imagination to visualize data as a concept. Metaphors play a similar role in our sensemaking practices, and the metaphors we use to understand data are of critical importance as, according to Star, 'power is about *whose* metaphor brings worlds together, and holds them there'¹⁸ and power in its many forms is a concern we will come to in Chapter 7. Euphemisms used by big tech companies to describe what they are then show what a slippery terrain we're on: so how do we get a grip on these fundamental

tenets of our era? Models like the Data Information Knowledge Wisdom (DIKW) hierarchy over-simplify data use by positioning data as a foundation from which we steadily build towards the keystone of wisdom. Leonelli instead advocates a relational approach to data and knowledge claims,¹⁹ which adopts a reflexivity that, in practice, acknowledges the roles of creators of data and knowledge and their ‘response-ability’²⁰ and requires a depth of engagement in the process that is not reflected in the norms of ethical guidelines. Leonelli also sees ethical box-ticking exercises as part of a system that upholds a simplistic view of data processing that denies context and the involvement of humans.²¹

As we shall see on our journey through the dimensions of data, knowledge and truth (and especially in Chapters 3 and 4), data practices have emerged through different traditions experiencing different digital disruptions. Ribes and Jackson’s investigation of the workings of the data archive 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.’²² What remain at repositories are the distilled products of these field sites; behind these centres lie an even more occluded set of activities that produce those data themselves. Extant research has not fully documented the tensions of the application of the ‘principles of information hiding’²³ to traditional archival practices, nor the extent to which existing metadata and practices across the sector already represent a big data approach to historical and cultural sources. *Archival thinking* and the organization of knowledge represent distinct cultures with their own values and norms while always anticipating change. Over seventy years ago, Broadfield described how knowledge 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.’²⁴ More recently, research has suggested that archivists are constantly changing and adapting their practices and systems up to and into the big data era.²⁵

Accordingly, the story of the organization of knowledge takes us into the development of ‘scientific’ practices synonymous with professionalization (and, as Susskind and Susskind have shown, that tee up the replacement of professions with machines²⁶), the growth of computer use from basic sorting functions to routine work to transforming practice, and the shift in decision making from the traditional gatekeepers of knowledge to data specialists testing the reach of

technologies developed for other purposes. While knowledge seekers may be accustomed to a weak grasp of the processes that facilitate their research, this black-boxed technology represents a double-bind and a further testament to objective data worship played out in every conceivable arena of society. Needless to say, how data is embraced or rejected is a matter of culturally inscribed tolerances, as big data and AI-driven approaches to knowledge creation and cultural shaping are both disruptors and creators of perceptions of authority, trustworthiness and credibility, as well as the knowledge claims evaluated in the light of this shifting perceptual capital.

In discussing big data in relation to archives, we were 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.

To understand what the digital turn really means for knowledge practices, it has been argued²⁷ 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'²⁸ is merely a useful discourse for those whose interests are served by the promulgation of an evangelical 'dataism'.²⁹ This phenomenon has parallels across society. For example, Williamson 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'.³⁰ 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.³¹ Berry³² draws on Fuller³³ 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.

These developments may be seen as playing out Technopoly's promise of efficiency,³⁴ which requires standardization, flattening of nuance, decontextualization and depoliticization of data by algorithms. The result of counting everything is that not everything *counts* – not only because developers avoid or fudge corner cases (as we will explore in Chapter 4) but also because the protective practices of practitioners resistant to data sharing can reinforce

invisibility. As we show in our discussion of our findings from the KPLEX study, those in roles of knowledge gatekeepers in conventional archival institutions are thoughtful in weighing up the pros and cons of increased exposure, taking a comprehensive view at least preferable to the blinkered perception of the developers who did not consider Facebook data to involve human subjects, for example (a particularly worrisome case study that will be discussed in more detail in Chapter 2). More recently, Facebook has been called upon to release data relating to its role in inciting genocide in Myanmar (a case discussed in Chapter 8), another 'bad practice' example we will return to in this book's conclusion. Representation through data is very often coterminous with the representation of people and identities, and attempts to obscure or protect these identities can be much more challenging than it may seem. At the macro-level, business models based on big data paradigms that assume appropriate safeguards can therefore be blindsided by their failure to prioritize cultural knowledge, including expertise in minority languages. The stakes, we assert, are too high when the pendulum of risk swings between genocide and the extinction of language. Big data and AI development trajectories are leaving countries with less internationally popular languages, such as Iceland, Latvia and many others, without access to the kinds of linguistic technologies that might enable them to balance the desire to preserve their cultural traditions with acting and competing in a globalized world. What *does* get shared and used is then also important (and potentially harmful) because of its eclipsing effect. Facebook drew attention to its tacit acknowledgement of this social fact by its actions in India, where the company demonstrated *computational thinking* by treating an entire country as an A/B test for its Internet.org project (and this in a country whose residents' lives are already subject to the social sorting behemoth of Aadhaar, a Kafkaesque tech-caste system). We explore such issues of power and the dangers of technology companies being allowed to behave as nation states in Chapter 6.

As the takeover of our knowledge commons by commercial interests becomes more and more an accepted norm, dangerous levels of skewing our information environment is made possible by the extremes of cannibalistic AI journalists³⁵ and data voids. Siri and Google Assistant are supposed to answer all our questions as long as they mainly relate to what the weather is currently doing, or obey an implicit or explicit script related to increasing or enhancing our role as consumers. Solutions will always be provided but not necessarily ones that answer our original question. The iconography of a magnifying glass is often used for search engines but this tool works in the opposite way to search algorithms, whose powerful position in narrowing horizons of information

discovery is discussed in Chapter 5. It is increasingly difficult for knowledge seekers to retain a sense of themselves outside of a paradigm that was heralded with the promise of democratization of knowledge, as only an elite few have the means to work with and control big data.

The above is not to say that we are advocates of calling up Bill Gates to switch the internet off. We are living in the After Google era and pointing at genies we might have wished to remain in their bottles is not our concern. What we do think is worthy of deeper consideration and analysis are the challenges society faces as big data logics collide with our established, valued and essential cultural norms and values. Though technology is itself also a part of culture, when it changes too fast or for too few, it becomes out of step with the identities of too many to support cohesion, displacing the old not in the sense of an organic assimilation, but as a rupture that creates fast-changing classes of epistemic ‘haves’ and ‘have nots’. Over the next seven chapters, we focus on sites of these tensions that may be seen as flashpoints for our relationships with data, knowledge and perhaps even with truth, to better understand the landscape of knowledge *solutions*, and the questions we should be asking.

Applying the KPLEX approach to these issues

These issues, and many others, will be expanded upon over the course of the next seven chapters, drawing together the existing basis of science and technology work with the culturally attuned perspective and results of the KPLEX project. The chapters will look at language, discourse and narrative (Chapter 2), sensemaking in the data-driven age (Chapter 3), the unintentional exposure and silencing of perspectives (Chapters 4–6), and how these trends drive power relations (Chapter 7). In the concluding chapter (Chapter 8), we will return to the high-level questions of culture and how big data development operates within epistemic, organizational and regional cultures, and the opportunities that might lie in this perspective to address some of the tensions and conflicts – individual, ethical and others – that the trouble with data has caused. The book incorporates a tension between the need to focus on culture and cultural practices being datafied or otherwise shaped by data practices, and our wider interest in knowledge creation, data and truth, and the many actors active in this ecosystem.

The book’s contribution to developing a language and method for applying humanistic methodologies and knowledge to the problem of big data as a knowledge technology will be laid out as follows:

Chapter 2 begins at the surface level, looking at how language and communication about and around big data research, both scientific and corporate, can be seen to mislead or, at best, fail in its intended purpose to build consensus around what is happening in the 'black box'. Chapter 3 introduces the social dimension of knowledge creation processes and explains two aspects of why big data can be understood as a disruptive force: (a) big data analyses in private corporations are mostly cut off from state-funded scientific communities and therefore from society as such and (b) where the mechanisms used in AI applications to analyse big data are unknown, knowledge creation processes become non-transparent.

The topic of Chapter 4 is resource discovery and search engines, such as Google, which have become the predominant method by which we access knowledge in our daily lives. However, the long tail of Web pages resultant from a keyword query will obfuscate unique research resources for the cultural heritage researcher. For the custodians of the original big data, the cultural heritage institutions, this new discovery mechanism also represents a challenge. They must not only ensure their users discover what is in their holdings but also understand how to expose the context and the tacit knowledge held by cultural heritage practitioners. As cultural heritage practice moves into the digital realm, Chapter 5 examines what other aspects may further lead to data becoming hidden from the historical record and the risks involved in making some data discoverable on-line. How the recently created data and research infrastructure can aid resource discovery for both institutions and their communities is also considered.

The dominant messages surrounding big data often neglect the complexities of people in the process, eliding human contributions to position the outputs of datafication centre stage. Chapter 6 looks at where datafication discourse may take our relationship with complex knowledge, and the historical record itself, by examining its impacts on the practices of archival practitioners and knowledge seekers. We show how big data precepts' crossover to the cultural realm is an example of how deference to human reasoning is being displaced, to the extent that AI might determine what we learn from the historical record, setting a troubling precedent.

Chapter 7 elaborates how big data approaches marginalize culture – here: language, cultural heritage, the culture of scientific knowledge creation – and explains how cultural commons have become transformed into big data which can be exploited by private companies, a development which has to be resisted. Finally, in Chapter 8, we return to the macro-level to open up the question of

how a cultural frame of reference might be further extended to serve the project of ensuring that datification processes remain in line with the public good, rather than favouring in such a monocular fashion the imperatives of engineers and the companies that employ them. We conclude with the proposal of one such mechanism, utilizing an intercultural studies approach to frame a new metaphor for the relationship between the users and creators of big data systems. With this provocation and the evidence that proceeds it, we hope to have succeeded in the goal of empowering both researchers and their subjects to make the development of big data approaches to knowledge creation more humane.

Notes

- 1 Henrich, Joseph, Steven J. Heine and Ara Norenzayan, 'The Weirdest People in the World?', *Behavioral and Brain Sciences*, 33/2–3 (2010), 61–83.
- 2 Drucker, Johanna, *Speclab* (Chicago: University of Chicago Press, 2009), 27.
- 3 'ENUMERATE Core Survey 2 (2014) Full Dataset – Excel Format – ENUMERATE Data Platform', https://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/ENUMERATE/raw%20data/ENUMERATE_datasheet%202017%20download.xlsx. See also http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/ENUMERATE/deliverables/DSI-2_Deliverable%20D4.4_Europeana_Report%20on%20ENUMERATE%20Core%20Survey%204.pdf.
- 4 'ENUMERATE Core Survey 2 (2014) Full Dataset – Excel Format – ENUMERATE Data Platform', https://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/ENUMERATE/raw%20data/ENUMERATE_datasheet%202017%20download.xlsx.
- 5 'Europeana Foundation | Europeana Pro', <https://pro.europeana.eu/about-us/foundation>.
- 6 *Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 Amending Directive 2003/98/EC on the Re-Use of Public Sector Information Text with EEA Relevance*, 175, 2013 <http://data.europa.eu/eli/dir/2013/37/oj/eng>.
- 7 European Commission, 'Draft European Open Science Agenda', 2016, https://ec.europa.eu/research/openscience/pdf/draft_european_open_science_agenda.pdf#view=fit&pagemo%20de=none.
- 8 Warwick, Claire, Melissa Terras, Paul Huntington, Nikoleta Pappa and Isabel Galina, *The LAIRAH Project: Log Analysis of Digital Resources in the Arts and Humanities Final Report to the Arts and Humanities Research Council* (School of Library, Archive, and Information Studies, September 2006), 60.

- 9 Edmond, Jennifer, 'How Scholars Read Now: When the Signal Is the Noise', *Digital Humanities Quarterly* 12/1 (2018).
- 10 Hall, Bud L., 'Beyond Epistemicide: Knowledge Democracy and Higher Education' (presented at the International Symposium on Higher Education in the Age of NeoLiberalism and Audit Cultures, University of Regina, 2015), https://dspace.library.uvic.ca/bitstream/handle/1828/6692/Hall_Budd_BeyondEpistemicide_2015.pdf?sequence=2&isAllowed=y.
- 11 As represented in work such as: Liu, Alan, 'Where Is Cultural Criticism in the Digital Humanities?', *Debates in the Digital Humanities*, <http://dhdebates.gc.cuny.edu/debates/text/20>; Hall, Gary, 'The Digital Humanities beyond Computing: A Postscript', *Culture Machine*, 12 (2011), <https://culturemachine.net/wp-content/uploads/2019/01/11-Digital-Humanities-441-894-1-PB.pdf>.
- 12 Edmond, Prof. dr. J. (Trinity College Dublin) (2018): Knowledge Complexity. DANS, <https://doi.org/10.17026/dans-xe6-hpm5>. All further references in this book to the KPLEX interviews will be referenced as follows (WPx INTy), where x refers to the workpackage that created the original interview and y refers to the interview number within that group.
- 13 Surgisphere, an American healthcare analytics company, provided datasets used for studies of Covid-19 that were published in *The Lancet* and *The New England Journal of Medicine* in May 2020. The analysis presented in *The Lancet* claimed to show that patients taking the anti-malaria drug hydroxychloroquine were more likely to die in hospital, prompting the World Health Organization to halt global trials of the drug for use in treating Covid-19 patients, before closer scrutiny revealed serious doubts about the data's provenance. The studies were retracted.
- 14 GCSE, A level and (Scottish) higher students in the UK did not sit exams in 2020 because schools were closed to most students in response to Covid-19. In England, teachers were asked to supply the official exam regulator, Ofqual, with an estimated grade and ranking for each pupil in all their subjects. These grades and rankings were then subject to an algorithm, which also factored in the school's previous performance, with the aim of allocating grades consistent with schools' prior achievements. The release of results prompted a public outcry, not least because the algorithm could be seen to 'downgrade' state school students, compared to those attending privately funded independent schools, a bias attributed to cohort size. As well as its implications for ideas of equal opportunities and social mobility, this system clearly undermined concerted individual effort. After his administration allowed some appeals resulting in upgrading, Prime Minister Boris Johnson told students: 'I am afraid your grades were almost derailed by a mutant algorithm.' The National Education Union (NEU) called Johnson's comments a 'brazen' attempt to 'idly shrug away a disaster that his own government created'. Coughlan, Sean,

- 26 August 2020, A-levels and GCSEs: Boris Johnson blames 'mutant algorithm' for exam fiasco <https://www.bbc.com/news/education-53923279>.
- 15 Leonelli, Sabine, 'What Difference Does Quantity Make? On the Epistemology of Big Data in Biology', *Big Data & Society*, 1/1 (2014), <https://doi.org/10.1177/2053951714534395>.
 - 16 Bernstein's (1973) term 'educational knowledge code' describes the set of principles that shape and define a particular discipline, demonstrating its distinct identity among other disciplines and setting boundaries for its curricula, appropriate forms of pedagogy and the means of evaluating teaching and learning.
 - 17 Haraway, Donna, 'Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective', *Feminist Studies*, 14/3 (1988), 575–99.
 - 18 Star, Susan Leigh, 'Power, Technology and the Phenomenology of Conventions: On Being Allergic to Onions', *The Sociological Review*, 38/1 (1990), 26–56.
 - 19 Leonelli, Sabina, 'What Counts as Scientific Data? A Relational Framework', *Philosophy of Science* 82/5 (1 December 2015), 810–21.
 - 20 Haraway, Donna, *When Species Meet* (Minneapolis: University of Minnesota Press, 2008).
 - 21 Leonelli, 'What Counts as Scientific Data? A Relational Framework', 810–21.
 - 22 Ribes, D. and S. J. Jackson, 'Data Bite Man: The Work of Sustaining a Long Term Study', in *'Raw Data' Is an Oxymoron* (Cambridge, MA: MIT Press, 2013), 147–66.
 - 23 McPherson, Tara, 'Why Are the Digital Humanities So White?' in *Debates in the Digital Humanities* (Minneapolis: University of Minnesota Press, 2012).
 - 24 Broadfield, A., *The Philosophy of Classification* (London: Grafton, 1946), 65–6.
 - 25 Daelen, Veerle Vanden, Jennifer Edmond, Petra Links, Mike Priddy, Linda Reijnhoudt, Václav Tollár et al., 'Sustainable Digital Publishing of Archival Catalogues of Twentieth-Century History Archives', 2015, <https://hal.inria.fr/hal-01281442>; Borgman, C.L., *Big Data, Little Data, No Data: Scholarship in the Networked World* (Cambridge, MA: MIT Press, 2015).
 - 26 Susskind, Richard and Daniel Susskind, *The Future of the Professions. How Technology Will Transform the Work of Human Experts* (Oxford: Oxford University Press, 2015).
 - 27 Bolin, G. and J. Andersson Schwarz, 'Heuristics of the Algorithm: Big Data, User Interpretation and Institutional Translation', *Big Data & Society*, 2/2 (2015), <https://cyberleninka.org/article/n/311316>; Kitchin, Rob, 'Big Data, New Epistemologies and Paradigm Shifts', *Big Data & Society*, 1/1 (2014), 2053951714528481.
 - 28 boyd, danah and Kate Crawford, 'Critical Questions for Big Data', *Information, Communication & Society*, 15/5 (2012), 662–79.
 - 29 van Dijck, Jose, 'Datafication, Dataism and Dataveillance: Big Data between Scientific Paradigm and Ideology', *Surveillance & Society*, 12/2 (2014), 197–208.

- 30 Williamson, Ben, 'Political Computational Thinking: Policy Networks, Digital Governance and "Learning to Code"', *Critical Policy Studies*, 10/1 (2016), 39–58.
- 31 Williamson, 'Political Computational Thinking: Policy Networks, Digital Governance and "Learning to Code"'.
- 32 Berry, David, 'The Computational Turn: Thinking about the Digital Humanities', *Culture Machine*, 12 (2011), 23.
- 33 Berry, D. M., 'Humanity: The Always Already – or Never to Be – Object of the Social Sciences?', in *The Social Sciences and Democracy* (London: Palgrave, 2010).
- 34 Postman, Neil, *Technopoly: The Surrender of Culture to Technology* (New York: Penguin Random House, 1993).
- 35 Waterson, Jim, 'Microsoft's Robot Editor Confuses Mixed-Race Little Mix Singers', *The Guardian* (9 June 2020), section Technology, <http://www.theguardian.com/technology/2020/jun/09/microsofts-robot-journalist-confused-by-mixed-race-little-mix-singers>.