Will Historians Ever Have Big Data?

Theoretical and Infrastructural Perspectives

Jennifer Edmond(✉)

Trinity College Dublin, Dublin, Ireland
edmondj@tcd.ie

Abstract. Digital history has spawned many great individual projects, and proven its value as both a methodology for the interrogation of sources and as a medium for the presentation and communication of research results. But the experiences of projects building research infrastructure for historical research raise the question of whether these methods can scale toward the realisation of historical 'big data,' or whether there are hindrances to this goal inherent in our current conceptualisation of the intersection between historical methods and computational ones. This paper discusses a number of the current barriers discovered by large-scale historical research infrastructure projects, including heterogeneous conceptions of what data is, hidden elements of data and the epistemics of humanities research. At the project level, these issues can be managed, but if digital history is to scale and grow to fit the infrastructural capability available to it, then a revisiting of some of the conceptual underpinnings of digital historical studies will be required.

Keywords: Cliometrics · Epistemics · Cultural computing · Digital humanities · Big data · Provenance · Authority

1 Introduction

> *"Less Guessing. More Knowing. Analytics, Delivered."*
> Accenture Advertisement, Dublin Airport, April 2012

> *"Analysing Big Data, That's the secret to living happily ever after."*
> Winton Global Investment Management Advertisement,
> London Underground, May 2015

As a society and as a research community, we seem to worship 'big' data. But like any other product of the human race, our datasets are socially and individually constructed, and prone to error and bias - indeed in many cases it is this very individuality of datasets that is the mark of their provenance, the knowledge organisation framework their creator applied, and an inherent part of their utility as a foundation for knowledge creation. Even when they are 'clean' to a 'gold standard' or 'open' to a 'five star' rating, datasets of any complexity, analogue or digital, remain objects that need to be viewed in the context of their creation.

B. Bozic et al. (Eds.): CHDDH 2016, IFIP AICT 482, pp. 91–105, 2016.
DOI: 10.1007/978-3-319-46224-0_9

So long as the complexities of data and their sources remains visible to the user, this 'human all too human' variation maintains its capacity to be a strength, rather than a weakness. But statistical and engineering imperatives have fostered approaches based on the assumption that the increasing scale of data is a hallmark of increased knowledge, of authority, perhaps even of a sort of 'truth.' To further increase the scale of integration of data into truly big data is to hide that complexity, cultural specificity and the social constructedness of data in a 'black box,' and to flatten the nuances inherent in it that may be essential to its interpretation. This flattening has been identified as one of the hallmarks of the engineer's methods, however. In her controversial study of the parallel histories of the development of the UNIX operating system and the American Civil Rights movement, Tara McPherson describes this in terms of two modes of professional seeing, the lenticular and the stereoscopic:

> The ridged coating on 3-D postcards is actually a lenticular lens, a structural device that makes simultaneously viewing the various images contained on one card nearly impossible. The viewer can rotate the card to see any single image, but the lens itself makes seeing the images together very difficult, even as it conjoins them at a structural level (i.e., within the same card). In the post–civil rights United States, the lenticular is a way of organizing the world. It structures representations but also epistemologies. It also serves to secure our understandings of race in very narrow registers, fixating on sameness or difference while forestalling connection and interrelation…We might contrast the lenticular postcard to that wildly popular artifact of the industrial era, the stereoscope card. The stereoscope melds two different images into an imagined whole, privileging the whole; the lenticular image partitions and divides, privileging fragmentation. A lenticular logic is a logic of the fragment or the chunk, a way of seeing the world as discrete modules or nodes, a mode that suppresses relation and context. As such, the lenticular also manages and controls complexity.
>
> There are clearly practical advantages of such [UNIX-style] structures for coding, but they also underscore a worldview in which a troublesome part might be discarded without disrupting the whole. [1]

The combination of the first problem discussed above - the dissociation of data sources from the people, institutions and conditions that created them - with the application of this 'lenticular' worldview can lead to a data environment where macro-level patterns, regardless of whether they are truly meaningful or based on misconceived alignments, begin to drive our processes of knowledge creation. But this 'epistemics of the algorithm' is further complicated by yet another elephant in the data warehouse: that a lot of rich, historical data isn't digitally available. By pushing toward ever greater integration of digital data, we risk leaving behind the core strengths and reinterpretability of historical archives, aboriginal folk cultures, of discourses of morality and ethics, and of artistic creation. The result is a claim to objectivity which only stands because the many small subjectivities behind it have been reduced to noise, hidden in the computational black box. These instances of the lenticular frame creeping in to the presentation of data are often unintentional: what one person sees in a historical source may be completely different from what another person does. It is impossible to ignore the fact, however, that the process of changing digital signals can also be intentional, as the creation of the “Ello” social networking platform as a direct response to Twitter’s licensing practices illustrates.

The inchoate signals, these sub- and counter-narratives, are of great interest to historians, as they open the door to new interpretations and new understandings of the significance of events that may have been overlooked or underplayed by previous generations of scholars. This is evidenced in part by the rise of approaches such as transnational history, which privileges research questions that transcend the traditional barriers of nation or language, reflecting both the relationship between scholars and their objects of study and the historian's particular sensitivity toward fragmentation and diversity [2, 3].

Identifying the currents flowing in and between dominant historical narratives in the digital age can be a challenge, however. Digital history has spawned many great individual projects, and proven its value as both a methodology for the interrogation of sources and as a medium for the presentation and communication of research results. The findings of teams building infrastructure for historical research raise the question of whether these methods can scale toward the realisation of historical ‘big data,’ however. Infrastructure projects are generally constrained by the larger frame of technical and end user practice in which they operate, and their experiences raise the question of whether there may be deep-seated barriers to the embedding of historical research into a fully computational paradigm inherent in our current conceptualisation of the intersection between historical methods and computational ones.

The experiences of the projects aligned with the European research infrastructure DARIAH-EU [4] imply that it will be a very long term and challenging process to reach the goal of a big data research environment for history. In particular, this is borne out by the efforts of the European Holocaust Research Infrastructure project (EHRI) [5] and the Collaborative European Digital Archival Research Infrastructure project (CENDARI) [6]. These parallel experiences were explored in a 2016 DARIAH workshop on “Open History,” which validated the impression of each of these teams that the problems of creating large-scale historical research infrastructure were far more fundamental than previously understood. [7, 8] What follows is an attempt to build upon these practical experiences to introduce parameters for a new conceptual framework able to underpin systems, collaborations and perspectives able to realise the potential of big data for historical research.

There are a few caveats that must be applied to this discussion. First of all, there are many substantial data development projects currently moving forward within the historical domain that are having a pronounced positive impact for scholars. Many of these are focusing on the development of linked open data resources to support historical studies, including projects such as the Pelagios Linked Ancient Geodata In Open Systems [9], the Seshat Global History Databank [10], Open Context [11], and the TRAME [12] /CENDARI projects’ linked efforts to bring key resources for medieval studies (shelf marks, authority files etc.) into a more easily deployed LOD framework. There are also many kinds of historian, and many components to the matrix of sources historians will use in their investigations. For the purposes of this discussion, I will focus largely on the needs and methods of the contemporary historian, whose period of interest would likely be somewhere in the 19th or 20th Century. For this user profile, the primary mode of knowledge creation is the direct or technologically mediated interaction with historic materials, usually (but not always) held in the national, regional or local archives of the country in which the events of interest took place. This particular mode of historical research presents specific challenges for the development of data-driven methodologies, as opposed to medieval historians (whose research objects tend to be more sparse and more likely to be digitised) or archaeologists (whose primary research sources are very often not the objects and sites of interest themselves, but the data gathered from those sites by the original excavation teams).

2 What Is 'Data?'

Before turning to the implications of big data for the study of history as a specific set of resources and activities, it is important to understand what is meant for the purposes this discussion by the term 'data.' A Google search on the term returns over 5.5 billion hits, but the fact that the term is so well embedded in modern discourse does not necessarily mean that there is a consensus as to what it means. Many definitions, even thoughtful scholarly ones, associate the term with a factual or objective stance, as if data were a naturally occurring phenomenon. [13] Scholars trained in the humanities know to query any such simple association, however: as Lyotard argued so cogently in his landmark work, The Postmodern Condition, "Scientific knowledge is a kind of discourse." [14] Equally so, we must recognise that the data that underlie the construction of scientific discourse are not fact, nor are they objective, nor can they be honestly aligned with terms such as 'signal' or 'stimulus,' or the quite visceral (but misleading) 'raw data.' To become data, phenomena must be captured in some form, by some agent, signal must be separated from noise, like must be organised against like, transformations occur. These organisational processes are human-determined or human-led, and therefore cannot be seen as wholly objective or neutral. In this light, it is perhaps more instructive to apply Rosenberg's functional conception that 'facts are ontological, evidence is epistemological [and] data is rhetorical.' [15] This statement results from an historical investigation of the terms defined here, in which he seeks to unpick the overlaps within this commonly interwoven field of related terms. "Data means—and has meant for a very long time — that which is given prior to argument. As a consequence, the meaning of data must always shift with argumentative strategy and context — and with the history of both." It is in this sense that the modern historian's data will be indicated throughout the discussion that follows, as the source material bearing evidence and witness to events of the past, objects that stand at the beginning of the individual historian's process of knowledge creation, but which have already been curated and indeed created by other individuals, other actors from the past or present. As such, we must understand the historian's data as constructed, but not necessarily for the purpose to which that historian will put it.

There is a further conceptual gulf to investigate between Rosenberg's 'data as rhetorical' and very many of Google's 5.5 billion instances of the term, however. Not all data is digital. This seems an almost too obvious proposition to be worth stating, but the increasing penetration into modern life of devices and services reliant upon the digital availability of data (and the market forces driving their sustained development) threatens to eclipse this simple fact. Unique and historically relevant data abounds in paper formats in the libraries and archives of the world. It exists in performance and works of art. It exists in our perceptions of the natural world and of our constructed societies and cultures. It is embedded in languages and religions, in folk practices and instinctual responses.

An interesting perspective on the possible cost of computational interventions into this world is relayed in Todd Presner's article on the "Ethics of the Algorithm." [16] This article presents a reflection and a critique of the processes employed by the Shoah Foundation Virtual History Archive, a landmark project assembling over 50,000 video testimonies of survivors of genocide, starting with the Holocaust. In order to make such a massive collection of video data searchable, the project developed a set of topical keywords, each of which was applied to the testimony at one minute intervals. This was a human-driven process, with a team of more than fifty individuals working for several years to carry out the initial indexing. But even this carefully thought-through methodology carried a price, and Presner very effectively describes the impact of this process:

> "The effect is to turn the narrative into data amenable to computational processing. Significantly, this process is exactly the opposite of what historians usually do, namely to create narratives from data by emplotting source material, evidence, and established facts into a narrative…what goes missing in the "pursued objectivity" of the database is narrativity itself: from the dialogical emplotment of the events in sentences, phrases, and words in response to the interviewer's questions, to the tone, rhythm, and cadence of the voice, to the physical gestures, emotive qualities, and even the face itself…Needless to say, databases can only accommodate unambiguous enumeration, clear attributes, and definitive data values; everything else is not in the database. The point here is not to build a bigger, better, more totalizing database but that database as a genre always reaches its limits precisely at the limits of the data collected (or extracted, or indexed, or variously marked up) and the relationships that govern these data. We need narrative to interpret, understand, and make sense of data."

While Presner's work is based upon a specific project with what might now be seen as an outdated technical structure, it is still, in a 'lenticular versus stereoscopic' sense, a good illustration of the kinds of emotionally charged issues, such as culture, religion, identity, belonging, trauma etc. where the current focus on building bigger databases and faster algorithms threatens to leave behind ethical, cultural and all-too-human nuance. In this example, we can see how the drive for big data is coming into conflict with the need for rich data.

A contrasting example of the potential for conflict between big data and rich data can be seen in the history of the Europeana Digital Library. The founding vision for Europeana, dating back to 2005, was an ambitious one: "…to establish a digital library that is a single, direct and multilingual access point to the European cultural heritage and that ensures a common access to Europe's libraries, archives and museums." [17] To deliver upon this vision, Europeana developed the Europeana Semantic Elements (ESE) metadata standard, a Dublin-Core based profile designed to minimise the effort required of institutions to make their content available through Europeana, and to enable the widespread cross-searching through this massive data set. In the end, ESE achieved its stated goal of enabling the federation of huge numbers of digital objects representing Europe's cultural heritage, but at the cost of much of the richness in the original data.

"A problem arises if not sufficient semantics are represented in such a common language and if a data provider aims at going beyond the limits of it in terms of expressiveness or granularity. Therefore, disadvantages of the ESE are that there are not many elements for describing an object, that it is not intended that these can be specialised and that it is not possible to model complex (hierarchical) objects, for example a monograph and its subparts." [17] ESE was replaced in Europeana with the new Europeana Data Model (EDM) starting in 2013, a migration process that continues to unfold, as Europeana seeks to harness linked open data and the capacities of its new model for richer description and more complex object hierarchies (crucial, for example, in the representation of archival records) to improve its functionality.

To create an approach to big data with higher potential for outward transfer and application within historical research, we start off mindful of Presner's conjecture that computational systems create data from narrative, while historians create narratives from data. The underlying assumption is that knowledge is derived from information which is derived from data, and there the process must begin. There is no path according to this model directly from knowledge to knowledge. However in the human world, the processes are more iterative and non-linear: knowledge can even create information, which can be manifested in data. The place of data in the knowledge creation process will therefore be a part, but by no means all, of what we will need to further investigate in order to truly facilitate big data approaches to history. It is, however, a fundamental requirement, allowing us to capture not just what the computational approach to data may flatten, but also the concomitant processes developed not just by humanists, but by humans, to preserve and make meaning from these noisy signals.

3 Three Challenges to Building Big Data Infrastructure for History

Once we accept the nature of data as socially constructed, we have a basis for understanding the challenges faced by current approaches to the creation of big data infrastructure for the study of history. This paper will discuss three of these in further detail. The first, **complexity of humanistic data**, has long been recognised. To increase the scale of activity, however, this element will need to be revisited and potentially assigned a new position in the developer's hierarchy of concerns. Second, we must develop more robust representational norms for **hidden data** implicated by the contents of a digital system. To not do so is to go against some of the most deep-seated impulses of the historical researcher, and to undermine the utility of digital methodologies for knowledge creation. Finally, there are great gains to be made in increasing our application and understanding of not just humanistic research activities (as captured in participatory or user-centred design processes), but also from digging more deeply into the cognitive and social elements of the **epistemics of historical and humanistic research**. Only through such an investigation can both the user and the reuse of data become more strongly conceptualised widely and applied.

3.1 Revisiting the Complexity of Humanistic Data

One of the foundational challenges of humanities research lies in the nature of its research objects: human beings, their languages, cultures and the records of their activities. Cultural signals (which, according to Manovich, constitute their own distinct level within new media alongside the computational [18]) can be ambiguous and are often conflicting and self-contradictory. This is true even in ‘low context’ cultures, where a greater cultural permeability is facilitated by explicitness in the communication and day-to-day deployment of cultural norms and practices, as inscribed most visibly in language, but also in personal interactions, in religious practices, and in artistic production.

In order to transform culture into something recognisable as data, its elements – as all phenomena that are being reduced to data – have to be classified, divided, and filed into taxonomies and ontologies. Even at their best, these processes rely on the ability to turn human communication into a set of rules for transactions, rules that are very often overturned or made more complex by the addition of fine nuances of tone, gesture, or reference. The stereoscopic world must be rendered lenticular, the narratives must become data. But the historian remembers or records what she discards in creating her interpretation, or at least remains aware that she discards. The computational system does not, or at least, does not generally do so in a manner transparent to the user. This lack of transparency presents a dilemma to historians considering digital methods and tools, reducing the scholar’s mastery of her methodological vehicle by which data has been turned into knowledge.

The tendency of technology is to turn its users into consumers rather than experts: for example, many of the most adept users of technical tools could not aspire to reconstructing the code behind them. But the black box is not an acceptable paradigm for research environments. A scholar needs to know when a result is underpinned by less robust algorithms or smaller bases for the statistical modelling, leading to less reliable results. For example, in large scale, multilingual environments (like Google Translate), variations in system reliability between languages and expressions is not communicated to the user. For historians to harness big data, the black boxes will need to become glass boxes – but how we contextualise this richer contextual information in a user-friendly fashion remains a challenge.

Investigating competing theories and definitions of data will only take us so far, as will superficial observations of our users. The CENDARI project deployed a suite of four different measures of the course of the project’s active development to harvest and integrate historians’ perspectives into the system development: participatory design sessions, prototyping on research questions, a trusted user group and weekly testing cycles. Each of these mechanisms uncovered further layers of activity and requirement (including an early facilitated discussion to agree what was meant from different perspectives by the term ‘data’). This process revealed that to understand how and why the data processing functions between computer scientists and historians differ, we need to dig more deeply into those processes, but also to develop a more robust definition of what the characteristics and qualities of data are from a humanistic/cultural perspective as well as from a computational perspective. For example, **provenance** is a key concept for historians and collections management professionals: indeed, a source 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 not to downplay the good work of teams like the W3C provenance working group, which has established a solid model for the representation of provenance. [19] It is merely to say that modelling of uncertainty and complexity under these protocols would be labour intensive at best, and impossibly convoluted at worst: in particular as the standard itself is not designed to model uncertainty (though possible extensions to make this possible have been proposed). [20] To give an example, let us consider the collection of papers of Roger Casement held in the County Clare, Ireland archives. Here is an excerpt from the description of the papers (already an anomaly among more traditional archival fonds):

> Personal papers relating to the Irish patriot, Roger Casement were kept under lock and key in Clare County Council's stores since the late 1960s. The papers were presented to the council by the late Ignatius M. Houlihan in July 1969. The Ennis solicitor had received them as a gift from "a member of one of the noble families of Europe." …The papers, mainly letters, cover the last two years of Casement's life before he was executed by the British for his role in smuggling arms into Ireland for the 1916 rising. The last letter on file is one from Casement, dated April 4, 1916, just 11 days before his departure for Ireland on a German U-boat, which landed him at Banna Strand in Co. Kerry on Good Friday, 1916.
>
> "I came across the papers during an inventory of the council's archives. At first, I did a double take, I wasn't expecting something so exciting. I instantly recognised the value of them and their importance for Clare and I was anxious to make them accessible as soon as possible," explained Ms. [Roisin] Berry [archivist]. "They date from Casement's arrival in Germany in 1914 to the very month he leaves Germany in 1916 on the under 19 bound for Ireland. The documents address a range of different subjects including the enlisting of Irishmen in the First World War, the appointment of an envoy from England to the Vatican, the Findlay affair, the work of Fr. Crotty in German prison camps, writing articles for the press, keeping a diary and the desire for peace. [21]

This excerpt (and it is only an excerpt) brings out a number of highly interesting examples of the potential complexity of historical sources. No less than three previous owners of the papers are referenced (one of which is only known for his or her status as a member of the aristocracy). Their place in Casement's life (and indeed his own place in Irish history) is explained, chronologically and in terms of his thematic interests. The material status of the collection is given, including the fact that it consists of 'mainly' (but not exclusively?) letters. A surprising anecdote is relayed regarding how the archive came to realise they held such a significant collection, which illustrates how the largely tacit knowledge of the archivist enabled their discovery and initial interpretation. This example is not an exceptional one. How is this level of uncertainty, irregularity and richness to be captured and integrated, without hiding it 'like with like' alongside archival runs with much less convoluted narratives of discovery? Who is to say what in this account is 'signal' and what 'noise'? Who can judge what critical pieces of information are still missing? These are perhaps more questions of "documentation" than "cataloguing" (to borrow Suzanne Briet's [22] canonical distinction between the two) but while Briet proposed that documentation approaches could be differentiated according to each discipline, the granularity she was proposing was far less detailed than anything that would be required for historical enquiry. Indeed, the focus of the documentation required would vary not only for each historian, but quite likely as well according to each of their individual research questions, a result of the historians' research and epistemic processes that greatly raises the bar for description within their digital resources.

Unfortunately, another key aspect of what historians seek in their data is **completeness**. In spite of the often fragmentary nature of analogue sources, digital sources are held by them to a higher standard, and expected to include all relevant material. This fact has been tested, and again and again, the same insight emerges: "Researchers are wary of digital resources that are either incomplete or highly-selective." [23] "One concern of humanities users … is the extent of the resource: whether the whole of the physical collection is digitized or not." [24] "Two key concerns for digital archives in general…are the desire to be: authoritative and of known quality [and] complete, or at least sampled in a well-controlled and well-documented manner." [25] This perception results from a somewhat outdated paradigm of the digital resource (that its only value is in the access it provides), and places a particular burden given the often hidden nature of many sources (discussed below).

A further key issue in the ecosystem is the relationship between **metadata** and the objects they represent, as well as their changing place in the research process: as reminders from a pre-digital age of physical catalogues; as the most common data to be found in digital systems of cultural data; as research objects that are seldom the focus of modern historical research in themselves; as structured data of a sort that is easy to aggregate; as a draw on the resources of the institutions that must create it; 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." [26] We will also, however, need to look into how emerging computational approaches, such as ultra large system approaches [27] and deep learning, may be disrupting the need for 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.

3.2 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 digitised [28]. Although great progress was expected by the respondent institutions in the near future, 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 also has to imagine that there would be large gaps in the availability of data from some countries compared to others (an assumption borne out by the experiences of research infrastructure projects).

Is this something that historians are unaware of? Of course not. Does it have the potential to effect the range of research questions that are proposed and pursued by modern historians. Absolutely. Modern historians often pride themselves on being "source-led" and characterise the process by which they define research questions as one of finding a "gap" in the current research landscape. Because digital data is more readily accessible, and can be browsed speculatively without the investment of travel to the source, they have the potential to lead (as the 'grand narratives' of history once did before them [29]) or at least incentivise certain kinds of research based on certain kinds of collections. The threat that our narratives of history and identity might thin out to become based on only the most visible sources, places and narratives is high. Source material that has not been digitised, and indeed may not even be represented in an openly accessible catalogue, 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 material is available online, and in particular that an increasing proportion of the most well-used and well-financed cultural collections are, means that the novice user of these collections will likely focus on what is visible, an allocation of attention that may or may not crystallise into a tacit assumption that what cannot be found does not exist. In the analogue age, this was less likely to happen, as collections would available only as objects physically contextualised with their complements: the materiality would be able to speak of the scale of collections, and extension into less well-trodden territory would require only an incremental increase in time or insight, rather than a potentially wasted research journey.

Sources are not only hidden from the aggregated, on-line view because they have not been digitised, however. Increasingly, users are becoming frustrated with digital silos. The current paradigm is not that a user visits a number of news or information sites, but that he channels his content through an intermediary, such as Facebook or Twitter. The increase in the use of APIs and other technologies (including personalisation 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 license. Most national, regional or local institutions hesitate to do the same, however, in part because of technical or resource barriers, but also to a great extent because they do not trust the intermediaries and reuse paradigms that are emerging. These institutions have developed over centuries to protect the provenance of items in their care, and to prevent their destruction or misuse. 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 is not an issue of copyright: it is one of trust and social contracts. It is also not an issue of putting all historical material online, or even indeed of ensuring it all is digitised: it is a challenge of ensuring that data can be used outside of the silos that were designed to hold them, and that what is not online can be clearly signposted alongside cognate collections. As complex as they may be, solving these particular problems is an essential requirement for transnational digital approaches to the study of the modern era to become possible, not to even think of their becoming widespread.

The following excerpt from one of the CENDARI project user scenarios (documented in the project's Domain Use Cases report [30]) provides an illustration of the challenges a transnational research question can pose in a dispersed source landscape based upon national silos.

> My project examines how the rural-urban divide shaped Habsburg Austrian society's experience of the war from about 1915 (when food and food shortages became increasingly politicized) and to what extent that divide shaped the course of the Habsburg Monarchy's political dissolution in the fall of 1918. I will focus on provinces with large multiethnic urban centers that experienced food crises: Lower Austria (Vienna), Bohemia (Prague), Moravia (Brno), the Littoral (Trieste), and Galicia (Krakow). … transcended the urban-rural divide—also grew sharper over the course of the war. I want to answer the following questions: How did the administration and realities of rationing vary between cities on the one hand, and between urban centers and the rural areas of their provinces on the other? How did food protests—and other grassroots demonstrations without party-political leadership—vary between these selected provincial capitals and within their largely rural provinces? To what extent were protesters' grievances cast in terms of urban-rural divides or in terms of other fault lines and antagonisms? How did inhabitants of these cities and their rural hinterlands experience and perceive the political dissolution of the monarchy in different ways, i.e. in terms of expectations and demands? To what extent did successor states —Austria, Czechoslovakia, Poland, Yugoslavia, and Italy—overcome, institutionalize, or exacerbate rural-urban divides?

This researcher's work covers four current national systems and at least as many languages. Because the work encompasses rural and urban contexts, it is likely that some of the required source material will be held in smaller regional or local archives (which usually have far inferior infrastructure to their flagship national equivalents). The work is looking at events, perceptions and interpretations that may not have been captured in the official records, and which indeed may only be measurable through proxy data or personal accounts. Even in the case of the successor states listed, two have since dissolved. This scholar is setting out on a rich transnational research trajectory, to be sure, but there will be very little support in the formal finding aids to assist in wayfinding or knowledge creation, and very little this individual will be able to do to progress such an ambitious project within the current landscape of digital resources, where countries such as Hungary are particularly poorly represented, in spite of the centrality of the legacy of the Austro-Hungarian empire for understanding the development of European structures and identities after that empire's fall.

3.3 Knowledge Organisation and Epistemics of 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.

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 that the individual researcher cannot independently answer. 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 Log Analysis of Digital Resources in the Arts and Humanities (or LAIRAH) project [31] pointed toward a number of key issues leading to a lack of reuse of digital data prepared by research projects. In particular, the lack of an early conceptualisation 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 organisation 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 organisational model was key to their epistemic process. As the yet-to-be published results of a research project [32] at Trinity College Dublin 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, quite hybrid, and already marked by the curatorial impulse that is preparing it to contribute to insight.

4 Conclusion

When you study human beings, your input data is already marked by the 'ownership,' intellectual or otherwise, of others. Managing this web of narratives and influences is one of the key talents of the humanistic researcher generally, and of the historian in particular, but it does also complicate their relationship to their data on a number of levels. In spite of the great progress technical frameworks and approaches within digital history have made in the past decade, much of this knowledge creation process remains either unrecognised or underutilised in the development of tools, services and approaches to support this field of research. At the project level, these issues can be managed, but if digital history is to scale and grow to fit the infrastructural capability available to it, then a revisiting of some of the conceptual underpinnings of digital historical studies will be required. A number of issues, more social or cultural than technical, will need to be addressed before this can happen, however. First, mechanisms must be formed for better collaboration and communication between computer science, information science and historians. This will involve not only interaction and dialogue, but also self-reflection. For example, until historians better understand their own data and epistemic processes, their dissatisfaction with current platforms can only be expressed and addressed in the most generic terms. On the other side, we should also be querying the imbalance in upskilling opportunities: there are many, many training programmes, summers schools, web resources and the like inviting humanists to learn programming skills, but where is the summer school introducing humanistic methods to computer scientists?

Second, we need to move beyond the mental model of mass aggregation for cultural data, and imagine instead systems that don't assume an endpoint where all material is digital. What would a hybrid system for historical research look like? Google Street View? The Internet of Things? An aircraft autopilot? How we think about and speak about our data and systems is important, as are the metaphors we use to describe what an ideal system would be like. These metaphors need to mature so that we can reimagine not just the goal of supporting digital history, but also the path that leads us there.

Finally, we need to develop systems that support trust. The content holders need to trust that new pathways for the use of their materials will not lead to the exploitation of individuals or of the resources themselves. Only at that point (a vision currently emerging under the rubric of ‘data fluidity’ [33]) will the social and technical systems underpinning historical research be able support the methodological trends (such as transnational history) and policy imperatives (such as open research data) that are the emerging norms. In addition, scholars need to learn to trust the systems. This is not just a matter of expecting them to mature in their understanding of the affordances and limitations of the underlying technologies, but of creating technologies that can balance lenticular and stereoscopic vision, encompassing and giving access to uncertainty and richness without sacrificing discoverability. The systems also need to make their own limitations and assumptions available to the user. An underlying ontology, metadata schema, search algorithm, or curation practice can greatly effect the applicability of a digital resource for a given research question, and the historian investigating that question must be empowered to query and alter these parameters.

None of these issues will be easy to address, but to realise the full potential of digital technologies and data for contemporary history, the community must surely ‘draw back so as to leap forward.’ The benefits of this process will not only support digital history, however, but also computational approaches in any number of areas, adding to the complexity, hybridity and extensibility of current systems.

Acknowledgements. This research is a partial result of work funded by the European Commission (CENDARI project, grant agreement 284432) and the Irish Research Council (Scholarly Primitives and Renewed Knowledge Led Exchanges). It also bears an intellectual debt to the members of the Knowledge Complexity (K-PLEX) project team, Mike Priddy, Jörg Lehman, Thomas Studolka and Rihards Kalnins, and to the two anonymous reviewers.

References

1. McPherson, T.: Why Are the Digital Humanities So White? or Thinking the Histories of Race and Computation. Debates Digit. Humanit. (2012)
2. Winter, J.: General introduction. In: Winter, J. (ed.) Cambridge History of the First World War, vol. 1, pp. 1–10. Cambridge University Press, Cambridge (2014)
3. Clavin, P.: Time, manner, place: writing modern european history in global, transnational and international contexts. Eur. Hist. Q. **40**(4), 624–640 (2010)
4. The Digital Research Infrastructure for Arts and Humanities. http://dariah.eu/
5. The European Holocaust Research Infrastructure. http://www.ehri-project.eu
6. The Collaborative European Digital Archival research Infrastructure. http://www.cendari.eu
7. Vanden Daelen, V., Edmond, J., Links, P., Priddy, M., Reijnhoudt, L., Tollar, V., Van Nispen, A.: Sustainable Digital Publishing of Archival Catalogues of Twentieth-Century History Archives. Final Report, Open History: Sustainable digital publishing of archival catalogues of twentieth-century history archives (2016)

8. Lehmann, J., Beneš, J., Bulatović, N., Edmond, J., Knežević, M., Morselli, F., Zamoiski, A.: The CENDARI White Book of Archives. Technical Report, CENDARI Project (2016)
9. Pelagios. http://commons.pelagios.org/
10. Turchin, P., Brennan, R., Currie, T., Feeney, K., Francois, P., Hoyer, D., Manning, J.G., Marciniak, A., Mullins, D., Palmisano, A., Peregrine, P., Turner, E., Whitehouse, H.: Seshat: the global history databank. Cliodynamics J. Quant. Hist. Cult. Evol. **6**(1) (2015)
11. Open Context. http://opencontext.org/
12. Text and Manuscript Transmission in Medieval Europe (TRAME). http://trame.fefonlus.it/trame/index.html
13. Rowley, J.: The wisdom hierarchy: representations of the DIKW hierarchy. J. Inf. Sci. **33**(2), 163–180 (2007)
14. Lyotard, J.F.: The Post-Modern Condition: A Report on Knowledge. Geoff Bennington and Brian Massumi (trans.) Manchester University Press, Manchester (1984)
15. Rosenberg, D.: Data Before The Fact. In Raw Data is and Oxymoron. In: Gitelman, L. (ed.), pp. 15–40 (2013)
16. Presner, T.: The Ethics of the Algorithm: Close and Distant Listening to the Shoah Foundation Visual History Archive. Forthcoming in: History Unlimited: Probing the Ethics of Holocaust Culture. Harvard University Press, Cambridge (2015)
17. Hennicke, S., Dröge, E., Trkulja, V., Iwanowa, J.: From ESE to EDM and Beyond: How Europeana Provides Access to its Cultural Heritage Objects. Informationsqualität und Wissensgenerierung. In: Proceedings der 3 DGI-Konferenz, pp. 129–140 (2014)
18. Manovich, L.: The Language of New Media. MIT Press, Cambridge (2002)
19. An Overview of the PROV Family of Documents. https://www.w3.org/TR/prov-overview/
20. De Nies, T., Coppens, S., Mannens, E., Van de Walle, R.: Modeling uncertain provenance and provenance of uncertainty in W3C PROV. In: Proceedings of the 22nd International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee (2013)
21. Casement documents are found in Clare. http://www.clarelibrary.ie/eolas/archives/casement_docs.htm
22. Briet, S.: What is documentation? In: Day, R.E., Martinet, L., Anghelescu, H.G.B. (trans.), Day, R.E., Martinet, L. (eds.) What is Documentation? English Translation of the Classic French Text. Scarecrow, Lanham, MD (2006)
23. Bulger, M.E., Meyer, E.T., De la Flor, G., Terras, M., Wyatt, S., Jirotka, M., Madsen, C.M.: Reinventing research? Information practices in the humanities. Research Information Network Report (2011)
24. Terzi, P.: Establishment of Trustworthiness in the Digitization Project ‘International Dunhuang Project. Masters Thesis, Swedish School of Library and Information Science (2015)
25. Dix, A., Cowgill, R., Bashford, C., McVeigh, S., Ridgewell, R: Authority and judgement in the digital archive. In: Proceedings of the 1st International Workshop on Digital Libraries for Musicology, pp. 1–8 (2014)
26. Drucker, J.: SpecLab. U of Chicago Press, Chicago (2010)
27. Edmond, J., Bulatovic, N., O’Connor, A.: The Taste of ‘Data Soup’ and the creation of a pipeline for transnational historical research. J. Jpn. Assoc. Digit. Humanit. **1**(1), 107–122 (2015)
28. Enumerate. http://www.enumerate.eu/en/statistics
29. Lyotard, J.F.: The Post-Modern Condition: A Report on Knowledge. Geoff Bennington and Brian Massumi (trans.) Manchester University Press, Manchester (1984)
30. CENDARI Project: Domain Use Cases, Technical report (2013)

31. Warwick, C., Terras, M., Huntington, P., Pappa, N., Galina, I.: Log Analysis of Internet Resources in the Arts and Humanities. Final Report (2006)
32. Edmond, J., O'Connor, A., Bagalkot, N.: Scholarly primitives and renewed knowledge led exchanges (SPARKLE), funded by the Irish Research Council
33. Romary, L., Mertens, M., Baillot, A.: Data fluidity in DARIAH – pushing the agenda forward. BIBLIOTHEK Forschung und Praxis, De Gruyter **39**(3), 350–357 (2016)