Edmond, Jennifer, Nicola Horsley, Jörg Lehmann, and Mike Priddy. "Please mind the gap: The problems of information voids in the knowledge discovery process." *The Trouble With Big Data: How Datafication Displaces Cultural Practices*. London,: Bloomsbury Academic, 2022. 63–87. *Bloomsbury Collections*. Web. 26 Apr. 2023. <<http://dx.doi.org/10.5040/9781350239654.ch-4>>.

Downloaded from Bloomsbury Collections, www.bloomsburycollections.com, 26 April 2023, 11:02 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.

Please mind the gap: The problems of information voids in the knowledge discovery process

We have outsourced our memory to online search engines, in particular Google, and just search for the ‘fact’ we cannot remember. Even the proprietary name Google, now one of the world’s biggest technology companies,^{1,2} has transmogrified into a verb: ‘Why don’t you google it?’ Web search engines have become embedded into our daily lives; we view information and knowledge through the lens of a search engine, generating a one-dimensional list-map of the knowledge landscape related to the keyword search query we have served to a ranking algorithm parsing a powerful, but by no means exhaustive, index of information sources. In over 70 per cent of cases,^{3,4} our process of seeking an information source is satisfied⁵ within the first page of results; however, for complex, culturally entangled research questions the most significant sources in the knowledge landscape are rarely near the top of the ranked query results. Not all research resources are accessible through search queries and even knowledge of their existence may not be available to the search engine user.

In many research disciplines where there exist well-known big data datasets which have been collaboratively created or are the output of sensors or other devices, there is little need to use search engines or AI agents for data discovery. However, in disciplines such as the humanities and the social sciences⁶ research data may come from disparate sources often held by more than one cultural heritage institution (CHI), accompanied by the kinds of rich contextual flows discussed in Chapter 2.

In this chapter we will discuss the challenges of gaps in the knowledge creation process caused by the process of locating research resources within the human record through search engines and how this is reshaping the response of the traditional keepers of cultural heritage data to the changing ways in which their communities of researchers wish to interact.

The dominance of search engines

The strength of the World Wide Web (Web) is in the hyperlink. It seems obvious now, as we can spend an inordinate amount of our time ‘clicking’ links in the hope that the next Web page of information delivered to our screen is the one that really answers the question we have. The Web’s success lies in its ability to satisfy with little risk,⁷ but does it do this, or do we spend too much time searching for a specific nugget of knowledge? More specifically, can a researcher investigating a complex cultural or historical issue actually find a specific piece of information or dataset held in a CHI before, or even without, turning up at their door? And, if she can find it, can she use it? Is its context clear? Is it complete, or its gaps visible? Is it in a language the researcher can read?

Tim Berners-Lee, in his CERN paper of 1989 ‘Information Management: A Proposal’ suggested that CERN could benefit from ‘a web of notes with links (like references) between them is far more useful than a fixed hierarchical system.’⁸ Indeed, since many researchers at CERN were on secondment, bound to return to their home institutions, a web of notes accessible from anywhere was what was needed to maintain research activities, and was what Berners-Lee & Robert Cailliau developed.

HyperText is a way to link and access information of various kinds as a web of nodes in which the user can browse at will. ... There is a potential large benefit from the integration of a variety of systems in a way which allows a user to follow links pointing from one piece of information to another one.⁹

Ted Nelson introduced the term ‘hypertext’ to express a corpus of ‘written or pictorial material’ which could not be ‘presented or represented on paper’ due to its complexity of interconnections. Furthermore, the hypertext may include summaries and maps of its content, and like any scholarly text it may have the addition of annotations and footnotes by scholars.¹⁰

Its purpose was to create techniques for handling personal file systems and manuscripts in progress. These two purposes are closely related and not sharply distinct. Many writers and research professionals have files or collections of notes which are tied to manuscripts in progress. Indeed, often personal files shade into manuscripts, and the assembly of textual notes becomes the writing of text without a sharp break.¹¹

We are only now approaching Nelson’s vision for scholarly communication,¹² via Berners-Lee & Cailliau’s World Wide Web, with the first examples of hypertext annotation services and research infrastructures that facilitate the discovery,

gathering and management of research resources. As can be observed with the Semantic Web,¹³ where information is more useful if it is expressly linked to other information, associating new knowledge with existing knowledge in a data infrastructure may present researchers with a transformative tool rather than the paginated metaphor of the Web.

The linking of snippets of knowledge, or as semantic assertions¹⁴ via a browsable (knowledge) graph,¹⁵ to form insight, understanding and new knowledge, can have both advantages (trust through association) and disadvantages (incorrect assertions). We see this data linking value occur in scholarly activities such as the use of nanopublications,¹⁶ and not-for-profit activities such as Wikidata¹⁷ and GeoNames.¹⁸

As the Web grew rapidly with so much information to discover and digest, all users faced information overload.¹⁹ Browsing presented little risk but cost in terms of time spent in fruitless foraging. However, browsing led to the discovery of serendipitous knowledge, and potentially more time wasted. In an environment of favourite site listings and printed directories,²⁰ search engines with large databases of keyworded URLs emerged as a method to attempt to organize knowledge through curated interfaces such as that of Yahoo! at the time. The need to access information accurately and quickly led to the rise of the Web search engine²¹ and in particular Google in 1998.²² At this time browsing the Web was as important as searching. Then, Google changed everything. The Google homepage still looks similar today as it always has; however, it was how Google collected its index and the page-ranking algorithm that made it the preeminent search engine and the *de facto* way we locate and access Web resources, eclipsing the variety and potentially diverse strengths of the kinds of approaches common to specialists in cultural heritage.

Ranking and the long tail problem

What makes Google different can, in part, be attributed to its PageRank algorithm²³ which ranked Web pages by an objective measure of link importance utilizing the topology of the interconnection of hyperlinks to calculate the ranking, relevance or quality of each Web page,²⁴ or authority in the graph of links created by PageSearch.²⁵ Unlike other search engines, at the time of Google's launch, it was able to produce search result lists with the most relevant and useful Web pages at the top.²⁶ Web users could still encounter problems in finding relevant information since when a simple keyword query is entered and

the resultant list of Web pages are ranked on their similarity to the keywords²⁷ which may not be well chosen. Kosala and Blockeel considered that there was a problem of low precision resulting from the irrelevance of many of the search results. Such low precision is part of the so-called long-tail problem of search queries where, for any general query, the majority of resultant Web page references listed have low rankings due to few interconnecting inbound hyperlinks. This is a phenomenon that is well known and a particular challenge for complex cultural research questions, when the research resources of interest are likely to be referenced uniquely (thus lowly ranked by PageRank) amongst tens or even hundreds of thousands of possible results listed. In 2012 Google started to roll out the use of a knowledge graph to aid the recognition of, and the discovery of, real world entities.²⁸ This knowledge graph is visually presented as a side panel with associated details about the entity and suggestions for browsing similar entities in the form of 'People also search for' links. However, the knowledge graph does not appear to provide additional information for entities which do not have an entry in a key resource such as Wikipedia or Google Books. Thus, even with accessing the Google knowledge graph, the vast wealth of results is overwhelming in terms of numbers but not quality. Even with refining a search query using the Google common search techniques,²⁹ a researcher is still required to hunt through considerable quantities of resultant Web pages.

Equally, for CHIs holding resources of relevance to scholars, it is difficult to promote their holdings via the likes of Google even if they have expertise in search engine optimization (SEO), which is examined further in Chapter 6. Academic researchers are not necessarily the largest demographic of users for all CHIs, though they are a community which take up a significant proportion of time for the institution's staff.³⁰ How an institution manages its migration to an external digital representation and communications is perhaps one of their biggest challenges currently. The browse metaphor³¹ may prove to be a better discovery strategy for unique resources (with few, if any, inbound references or links) that may only appear deep in the long-tail. This is a strategy utilized by many research infrastructures and particularly by those who serve a specific research community. It is quite easy for research resources to become unintentionally hidden and this issue became the subject of one of the investigations carried out in the KPLEX project.

Interviews conducted with archivists, librarians and others working in CHIs revealed concerns and fears regarding the dominance of Google in the discovery of resources the CHIs hold as well as other challenges the cultural heritage sector faces from technology, the internet and big data.

Cultural heritage institutions: The original custodians of big data

Each CHI is like a jigsaw piece in that they are uniquely shaped by their own identity, they are what they keep and they keep what they are,³² and together they create an enduring cultural knowledge portrait, upon which researchers may draw. As such, CHIs are engines of knowledge creation.³³ Each institution has its own particular remit and mission, which is reflected in its acquisition policy, its user communities and its activities.

Technology has already made a considerable impact on CHIs, as the KPLEX study's survey of practitioners shows (see Figure 1), although it could

Figure 1 Types of holdings worked with by survey respondents.

be regarded that the cultural heritage sector is one of the last sectors in the knowledge and creative economy to resist wide-scale transformative digital change. It may be regarded that cultural heritage institutes are the original custodians of big data, just that the data is contained in artefacts and on paper rather than in bits and bytes, resisting the transformation due to the scale of holdings, which archives measure in kilometres of shelving holding boxes and folders and which for museum collections can run into hundreds of thousands if not millions of objects³⁴ with only a small proportion on display. The challenge of digitizing this wealth of material may never be completed, due to the lack of resources, so it will never truly become big digital data that can be manipulated and explored by algorithmic means. However, the big data era provides a case study of how cultural heritage practices may be changed by influences external to the institution. Understanding these influences is essential to an appreciation of complexity in the knowledge creation process.

CHIs provide expert services as knowledge gatekeepers

Cultural heritage practitioners' role in promoting the use of the holdings they acquire begins with their catalogue descriptions of the items. This process is recognized as being shaped by practitioners' personal backgrounds and their institutional cultures, as well as the power dynamics, of geography, class and gender, which govern the construction of meaning more broadly. However, practitioners' representation of the historical record requires them to be 'neutral' intermediaries between users and information.

A fundamental part of cultural heritage practitioners' practice is helping researchers with research problems and methods. Each of the participants' institutions provided help for researchers as part of their daily operations.

If people come in with a specific research question, we can point them in the right direction and work on a specific source ... the difficulties we have now are that ... most of our collections are hidden in the sense that the collection description is not yet online ... Once they email, we usually start a dialogue ... because a research question in itself might not be specific enough to determine which collection would be most appropriate. ... We try to give, I guess, custom-made service to guide researchers one-on-one in their research, and once they come here, it's very labour-intensive.

(WP3 INT3)

The depth of interaction between cultural heritage practitioners and researchers, with practitioners being central to the discovery of resources, is essential in helping to shape the research process itself, as they guide researchers through the complexity of the resources and knowledge held.

Clearly, what a CHI can provide to aid a visiting researcher is considerably more than the query–response mode of a Google search remotely. However, in particular, uses of archival data were found to be changing as a result of the increased online visibility of descriptive metadata, research artefacts themselves and/or their underlying data. Researchers were approaching practitioners with more refined questions:

They are even coming to the reading room with [resource identifiers³⁵] to see documents. So, it's a new approach, because ten years ago, people would have to come to the reading room and ask if we have some information on a particular subject, and so on. And now, they come with a list of documents that they want to see.

(WP3 INT2)

Digital methods of communication are playing an important role in improving the discoverability of complex cultural heritage research resources, increasing both the speed of direct communication and the numbers of individuals supported, with a wider geographic dispersal. Communicating with researchers using institutions' own websites and those of research infrastructure portals was a significant channel for the regular dissemination of information (see Figure 2). However, only 39 per cent of respondents to the survey considered that they had a significant amount of their holdings described online with only 9 per cent stating that 100 per cent of collection metadata as available online (see Figure 3).

Practitioners understood that their role will change with any move to providing knowledge online:

A description [online] can never be neutral, but then researchers can read into them and make up their own minds. If they go into dialogue with us, my colleague and I, we have a certain way of thinking because we know the materials in a certain way. If we eliminate the staff member element in the equation, we will probably get different results. There's a lot to do, we won't be eliminated, but I think our role will also change once we put the descriptions online.

(WP3 INT3)

Figure 2 Methods of communicating information about collections to researchers reported by survey respondents.

Figure 3 Survey respondents' percentage of the information (metadata) describing their collections is available online to the general user.

Content versus context

A significant amount of a curators' energy goes into creating descriptions, and in the case of archives the collection level description takes priority, containing much of the contextual information likely to pique researchers' interest:³⁶

It's often the collection description that can trigger researchers to take a look. The more information the better, but it's also time-consuming. It takes me about ... if it's a big collection, it can take up to two or three days to make a collection description. ... We try to put the context into the collection description. That's one of the necessities.

(WP3 INT3)

The context where a document comes from is very important for interpreting it. And also, if you just draw out some document from this and there and so on you don't have a wide view of the whole, you just have this singular document and it can be misleading ... That's also why we try to make out the provenance of our holdings or write something about it. We not only describe it but we also write about where it comes from, how it came to us, and how it came into existence. Was it a collection of a researcher or is it the documents of an organization?

(WP3 INT1)

Description therefore plays a fundamental role in researchers' discovery of material valuable for their research. A consequence of this researcher need for discovery is that many of the early research infrastructures concentrated on aggregating and republishing descriptions of research resources rather than aiding and encouraging the 'publishing' of digitized versions of the *objects themselves*. For researchers investigating archival materials, context can be important in assessing the trustworthiness of the source; how the documents were gathered and by whom, why and when.

Curation goes beyond the basic description of an object and its context with other objects. An expert curator may add knowledge of individual's name changes (due to marriage or migration), variation in spelling and translation not found in a document, or provenance in terms of the journey taken by the object. General search engines cannot second-guess missing knowledge and discovery would not be possible without the expertise. Furthermore, a curator with subject expertise may identify additional resources of which the researcher had no knowledge and may not be identified through the finding aids. A finely crafted representation of context may provide the foundations of sound research

but researchers' experience and habits of exploring archival resources could be considered significant:

The [collection] description is so important and has to be standardized so that you get the same information from every collection description. We've discussed building a hierarchy in the portal website to show researchers where a specific document is physically, but ... it's not that important anymore. Once you know the context of a collection and you know in which folder the document is, that's often all the context they need.

(WP3 INT3)

CHIs are trying to make their holdings more accessible through various digital and non-digital means (see Figure 2); however, when only 9 per cent have their entire catalogue available online (see Figure 3), then significant quantities of potentially valuable resources are hidden from search engines and even for visitors to the institutional websites. Therefore, archives should consider changing their core practices to make their cultural heritage holdings more accessible:

I believe there will be a need to think about ... different strategies to make stuff even more easy and accessible but which will have to be ... outside the classical description [but rather] narrative-based introduction to collections ... making people a bit curious ... But even at best, we try our best, they will not be, I mean, completely usable by anybody ... there will be a limit to what we can explain by our type of description.

(WP3 INT6)

The challenge for institutions, which use a hierarchy of description for their holdings, is that users are being hyper-transported into the bowels of the archive by a search engine with little descriptive context at the deeper levels of catalogue description, rather than experiencing the more atmospheric and contemplative guided journey down the carefully custom-wrought ladders of the collection descriptions. This practitioner's concept of 'layers of documentation with pop-up explanations' (WP3 INT6) requires additional effort on the part of the institutions; therefore, they must prioritize what they are able to offer to different user groups and how those users work with their material and metadata: 'one size fits all' is cheaper and quicker to build. There is a tension between the traditionalists and those who appreciate that the approach to resource discovery is evolving:

We have archivists here, really traditional, who think that that the context is the only thing that users need. And so, they must have access to the archives with the logic of an archivist and with the logic of the ... hierarchy ... others say, no, people are researching with keywords, and they're searching for specific documents and they don't need to have more information about the context and they don't understand the way that archivists are organising archive documents, so that it's not necessary to keep the context. We think here that a good way to work is between both options, so to develop a tool which gives the opportunity to make a search with keywords, and maybe to make a more specific search.

(WP3 INT4)

Researchers needed to be conscious of their choice of approach and be skilled in carrying it out. In practice, even where a dual approach was adopted, the balance was seen to have swung in favour of the key word search:

For historians, for example, it's really important, but the fact is that the problem is that now even researchers, even historians are making their search with keywords and not with the logic of archives.

(WP3 INT4)

Consequences for users were, however, largely framed in terms of the benefits of discoverability and independence that came with digital dissemination:

But generally, it made a big difference with our database on the internet. People have different requests, more concise. They don't ask, do you have anything about this? They ask, can I have access to this and that? ... They will search on their own and we have less work on this part.

(WP3 INT1)

The user engagement functions of the CHIs are changing, but this may mean that the tacit knowledge of the gatekeeping practitioners is, at best, underutilized and at worst lost as staff retire or are redeployed. Staffing levels in relation to researcher enquiries can be an everyday challenge:

We try to [engage with researchers] but we have a limited staff. So, in the local depot or here [in the national archive], you can ask questions. 'I need that', 'can you help me with that?', and so we talk with the persons and we try to help. [We suggest a] source can be helpful ... When you have a problem with our digital archives, you'll have to fill in a form and maybe one half-hour per day I respond to these questions, but the responses are very standardized because we can't go

in. We don't have the staff to go into all the difficulties of their questions, so you give general directions.

(WP3 INT7)

Researchers' experience of this archive was less personalized as a result of the high ratio of users to practitioners, and although they might be engaged in a dialogue with an archivist, it was likely to be of a more instrumental nature, with a requirement on the researcher to stimulate and steer the process. Therefore, in this case of an overstretched archive, it would be more efficient and more fruitful if the researcher first engaged via online searches of the institution's catalogue.

CHIs have been cataloguing their holdings and giving context, for their entire existence, and this has evolved into institutional specific methodologies often based upon national requirements and standards which have transfigured into digital versions maintaining the vagaries of past practice.

Metadata was seen by practitioners as one of the vehicles for conveying context to the user but using systems that were fit-for-purpose and enabling the user to travel between perspectives also played a role:

I think generally it's useful if in the way you sort of structure your information you can capture relationships between objects and even things that aren't objects ... Neither our systems nor our metadata is sort of properly set up to [provide this context]. So, in that sense I see less value in just describing an item on its own and more value in trying to in the way collections are put together, made available, to sort of build some of those links or make them more visible.

(WP3 INT10)

External influences were cited as the catalyst for adopting greater standardization of metadata, as this account illustrates:

Before we introduced the international standard archival descriptions, we used a customized metadata schema. When the museum was created in 1994, the archival description was not meant as an archival description. The staff members were collecting objects – photos, documents – to make the permanent exhibition. They created a metadata schema especially focusing on what is the topic of this photo or this document, where could we put it in the museum, what is the keyword so we can quickly find it to put it in the layout of certain panels. When we made a transition in 2012, we did a mapping to see which fields were in the old schema and we tried to combine fields to fit the international standard, and that actually worked.

(WP3 INT3)

Standardization is challenging for all institutions as it not only costs time, effort and money, all of which are in short supply in CHIs,³⁷ but also exposes previous mistakes in methodologies utilized and variation in practitioner practice.

Spacelessness, placelessness and hypertravel

Most CHIs have a situated production (creation or holding) and consumption (observation) relationship with the population³⁸ that funds their activities and visits their physical location(s), be it at a national or local level. Their collection policy may often reflect this relationship, and the manner in which the institution was established. A national archive will hold documents of state and others of national significance, whereas a local museum may hold objects which may once have belonged to local dignitaries and luminaries, local archaeological finds, or related to local industries. Topic-specific CHIs will also usually have a national or local historic place-centred remit. Clearly there are examples where the institution's holdings are not, by and large, related to the place, in which they are located – famous examples being the Pergamon Museum³⁹ and the British Museum,⁴⁰ where collections were built upon the amassing of items by their respective nationals. Archive-museums such as Yad Vashem,⁴¹ located in Israel, have a global subject-specific collection policy which includes the copying of archival documents on the Holocaust from other institutions.

This traditionally location-centric nature of CHIs' activities supporting researchers has been greatly altered by the Web, the internet and faster new forms of telecommunications available. The notion of distance is rendered meaningless by telemediation, or rather it does 'significantly disrupt the spatial logic of modernist societies'.⁴² Moreover, Dodge and Kitchin argue that 'time is becoming *the* crucial dimension of who is accessible, rather than geographic location'.⁴³

With financial resources and available time becoming a limiting factor for researchers, travelling to CHIs can be targeted and made meaningful by utilizing the institution's online catalogue. The initial discovery phase of the research lifecycle may be made more efficient with specific judicious questions asked by the researcher after accessing the online catalogue, rather than a 'general question like, do you have something about Holocaust?' (WP3 INT8) first and seeking guidance from the archivists as to the possibilities for narrowing their exploration.

While the research process still involves a dialogue between practitioners and researchers, online catalogues have shifted the balance of control towards the user. For smaller institutions that had not previously enjoyed exposure to a wide audience, datafication of finding aids and online catalogues has expanded the proportion of material used:

Because before people ask only for the same documents and the same collections. Thanks to the digital library and the publication of the inventories we recorded an interest for other collections that before [had] no requests, not requested or very rarely requested. So, the digital library is a very good ... vehicle to make known all that we hold.

(WP3 INT8)

It may be argued that with the placelessness of the online resources, both catalogues and digitized collections, researchers, and in particular early career researchers, are missing the valuable experience of interacting with the knowledge of the cultural heritage practitioners, located with the research resources, and the practitioner's support in developing the research question. Communication tools, such as video conferencing, may be considered to increase the physical separation between the researcher and the practitioner and hence the placelessness of the interaction; however, they maintain the connectedness in a placeless environment. Nevertheless, differences in time zones may still affect accessibility to real-time services and resources.

Dodge and Kitchin question whether that placelessness represents the diminishing of an authentic identity of place and ask 'does cyberspace help render geographic space placeless?' They have 'little doubt that new places, and new spatialities are being formed online'.⁴⁴ However, the generated one-dimensional map of results from a Google search is entirely placeless; it does not represent a geographical place or for that matter a persistent virtual place. It could be contended that it does not represent the Google space or place. The context of the links, to collection holding institutions and others, is entirely removed from what is presented, the placeness and trustworthiness of the linked Web page are all but removed, with only the page title, additional URI⁴⁵ and a few words around the keywords found being visible to the searcher. The provenance and reliability of the potential research resource are not present and as such the researcher must approach the resource with caution, making further enquiries to establish the value of the resource knowledge.

The relatively low cost, spaceless nature of the Web provides the opportunity for cultural heritage institutes to open up their considerably larger holdings

than those normally on display to the public. With so much virtual real estate the challenge comes with the user experience design to ensure that the online experience matches or even betters that of physically visiting the establishment.

As stated earlier, there is little risk to the researcher in selecting any link which does not immediately provide the knowledge required; nevertheless, there is a risk for the CHI in that they may be overlooked. Moreover, the researcher may, when clicking the link, have been hyper-transported⁴⁶ into the depths of the institution's holdings online catalogue without any context at the deeper levels and therefore may leave as rapidly as they arrived, rather than exploring the collection description further. The CHI must be very savvy, not just in how they describe their holdings to be search engine compatible, but also how they present their place and services online – 'layers of documentation with pop-up explanations' (WP3 INT6) – to ensure the individual user's experience of their virtual place is the same, or if not better, than experiencing their physical place, although 'one size fits all' is cheaper and quicker to build. There is a tension between the traditionalists and those who appreciate that the approach to resource discovery is evolving:

We have archivists here, really traditional, who think that that the context is the only thing that users need. And so, they must have access to the archives with the logic of an archivist and with the logic of the ... hierarchy ... And so, others say, no, people are researching with keywords, and they're searching for specific documents and they don't need to have more information about the context and they don't understand the way that archivists are organising archive documents, so that it's not necessary to keep the context. We think here that a good way to work is between both options, so to develop a tool which gives the opportunity to make a search with keywords, and maybe to make a more specific search.

(WP3 INT4)

Google may be perceived as a threat to CHIs due to its cataloguing of the institution's entire Web presence, but it is also presenting the institution's collections in a placeless frame of reference, and as this practitioner states, missing context and other useful resources:

I never use the search engine when I want to do research. Because in the end you have this, and you see things, or maybe I can use that. And when you go the straight way, and even when the search is still methodological, then you miss things, because you don't look for other sources, which can be important.

(WP3 INT7)

Google as a threat

In the interviews conducted, Google was consistently referred to as a 'knowledge sharing platform' that represented a threat to greater openness. For instance, the traditional hierarchical structure of archival collections, through which contextual connections could be traced, was reported to be losing significance as it was undermined by keyword searching through Google and the like:

People are so adapted to the Google search that they don't even know anymore that there is a different way of searching. Therefore, we also chose to put in keywords and categories that allow researchers to browse through collections based on keywords and categories ... That can give you a very interesting perspective as well because it can show you connections between collections ... and it can help you find more proof for the point you want to make.

(WP3 INT3)

There was broad acknowledgement that the changing use of collections disrupted some of the fundamental tenets of archival cultural heritage preservation practice both through Google's presence as a rival source of knowledge and in the ubiquity of search engine use infiltrating archival research methods. Practitioners' support for integrating search engines into the research process was at least partly motivated by a concern that the satisficing return of a Google result might attain a hegemonic position that ultimately marginalized the role of the archive in knowledge creation. Many practitioners expressed caution about the 'quick wins' of the keyword search, favouring an understanding of resources and context:

You have to get to know your material for good research, and we can aid people with that, [advising them] to look into these sources and these sources, but [even] I learn things. So, I think context is that important ... and one of the main problems is that the new historians, the new school who use the digital tools, they don't have the feeling anymore with the context, the methodology of searching. They want quick wins. And you can't do decent research I think with quick wins.

(WP3 INT7)

The support given by the cultural heritage practitioners is naturally seen as a key component of the services they offer to researchers. Loss of this expertise will diminish the value of the research resources and make their discovery difficult.

The commercial nature of Google, despite the impartial appearance of a set of search results, troubled this librarian:

If Google can do better at discovery than we can then I mean by all means use Google. However, these organizations have commercial interest that we don't necessarily share and there're various biases built into neutral-like looking technologies that will probably skew the perception of what knowledge is and introduce all sorts of bias into it. Even though people believe they see everything, they might see even less than before because they're only being shown the things that the algorithm believes they want to see.

(WP3 INT10)

Understanding and being heedful of the inherent biases, which may sometimes be the commercial imperative of the business, is an approach which both researchers and CHIs must maintain when complex knowledge is being sought; important resources may become hidden by the algorithms. Although this participant was keen to 'jump Google' (WP3 INT8) and become the primary source for their holdings, so that 'Googling it' is not the researchers first thought to find relevant resources:

To be stronger than Google ... I think this is the challenge for that kind of project. To be very attractive for people and push them to jump Google and go directly to your portal.

(WP3 INT8)

Benefits of search engines and digital cultural heritage

Google has been compared to Johannes Guttenberg's mechanical printing press making information easier to access so that more people could benefit (Kane, 2009).⁴⁷ It could be argued that Google et al. have considerably changed the research process in terms of resource discovery for both researchers and cultural heritage practitioners involved in developing new systems.

And so, the aim is to give the opportunity to make the research by keywords, also by the logic of archives if necessary, with the context, but first of all with the keywords because even researchers are doing their search now by keywords. But when you find a document, then you can see the documents in its context, in its archival context, that's the aim and that's the idea that we see for the new tool.

(WP3 INT4)

The well-known interface of Google and its indexing ability may be considered to be preferable for researchers compared to an unknown institutional search engine (or browse-based system); however in both cases, a researcher cannot be certain that all the 'relevant' resources are returned for a specific search query.

Collections, which were purely digital, offered new possibilities through their fluid nature:

[There] is a notable difference from my archival colleagues, who are all about the original order in terms of preserving the paper copies because you can't just reorganise a box of papers without losing something, whereas in the digital realm, it's a bit easier to provide different views upon things without making irreversible changes; not always but ...

(WP3 INT5)

Although many CHIs may have run digitising projects for specific collections, or may scan documents upon request, digitizing everything was widely regarded as impossible.⁴⁸ Moreover, digitization does not necessarily convert the object into usable data so the incorporation of complex knowledge into future research using digital methods may be unpredictable.

Research methods drawing on digitization were also not static and interview participants saw a need to respond to evolution within the digital revolution. Advanced search methods are becoming less prominent due to advances in indexing and the use of natural language queries. This is not to say that the implementation of advanced search functionality in specialist research-orientated services will not be utilized if, when learnt, provide accurate speedy results. One archivist felt that 'the difference and the real evolution will come from the quality', in that, once the 'fashion' for launching digital tools had lost its novelty, only the best ideas would attract attention and investment (WP3 INT8). This could be considered an expensive proposition requiring many iterations to evolve to meet disparate needs and unlikely to attract investment to 'Cinderella' research domains and resource-strapped CHIs.

Beyond the keyword

The digital revolution has brought about the exchange of the conventional logic of the cultural heritage practice, and in particular archival practice, for the new logic of the keyword search which is a sea change for archival institutions, as archivists' knowledge of the deep connections of hierarchical context is being

circumvented and unrepresented in the generalized search engines which are so often the first port of call. At present digital discovery is still supplemental to the value of enquiry with a cultural heritage practitioner – increasing the speed and reducing the costs of the initial discovery phase of the research lifecycle for many researchers. How long this situation will last is uncertain, but it will be adapted and the knowledge creation process transformed. The tacit knowledge about the institutional holdings of the cultural heritage practitioner is essential to the knowledge creation process whilst research materials may remain ‘hidden’ from the digital realm, not being described digitally or digitized for access. Researchers who arrive at an institution’s reading room with a list of identifiers may at first glance appear efficient; it does still bypass the tacit knowledge of the practitioner and institution. A more holistic approach is therefore advocated once institutions have settled into digital practices and started to think more strategically about how they might serve the specific needs and aims of the research community more precisely.

If data are the new oil (a misleading metaphor discussed in Chapter 2), then like oil its profits are not evenly distributed with only some benefitting from its intrinsic value. The vast majority of data preserved in CHIs is not held in bits and bytes.⁴⁹ Thus, the likelihood that all of the produce of millennia of human endeavour being digitally accessible can be considered to be negligible; it will remain analogue in form for the foreseeable future. The use of computational big data techniques may only be applied in limited and specific ways to data which has most likely been created and prepared specifically for the research methodology employed, consequently never truly big data on the scale seen in other research domains. Reuse may be limited due to the effort input during data preparation leading to an unwillingness to share the dataset with other researchers.

It is important to note the technical differences between Google’s search engine and the ones provided by CHIs. Google uses cookies and archives the search history of its users; thus, each ‘Google search query produces a wake of collateral data such as the number and pattern of search terms, how a query is phrased, spelling, punctuation, dwell times, click patterns, and location.’⁵⁰ These collateral data go far beyond what a user usually associates with her search activities – for example, ‘websites visited, psychographics, browsing activities, and information about previous advertisements that the user has been shown, selected, and/or made purchases after viewing’.⁵¹ This data surplus enables a prediction of the future behaviour of the users, which is then marketed to Google’s customers and which forms the basis for the placement of advertisement. Search engines

provided by CHIs or research infrastructures, on the other hand, are based on controlled vocabulary and knowledge graphs and not on data collected via cookies. In the case of libraries, recommender systems may be in place, but they also do not take previous searches by a researcher into account, but use established indexes or available linked open data to produce results relevant to the researchers. While the differing technical set-up may lead to heightened expectations on the side of the researchers accustomed to Google searches, and therefore to frustrations with the search results presented by an archives' search engine, it becomes clear that no archive will ever provide a search engine as sophisticated as Google's and that archives' technical facilities will never match cultural heritage practitioners' capabilities to answer to the research questions of their users.

Digitized images, or scans, of documents may solve issues of distribution for the CHI but for those researchers wishing to perform, for example, computational linguistic analysis they are less than ideal, requiring further transformation. Whereas a digital representation of a historical photograph may be of sufficient quality for another researcher to reference in their monograph. As one practitioner put it, 'nobody is, like, proud of his huge digitization project' (WP3 INT6) but the essential question for CHIs is what should they do with the digital materials, where any strategy must include a policy where the 'tendency will be towards openness' (WP3 INT6) and sharing digitally will bring together collections and resources, which may be physically very dispersed to create new knowledge and insight. Thus, sharing may facilitate not only more use, but also the recontextualization of materials for example Frankl et al.,⁵² and in archival research guides such as Frankl and Schellenbacher⁵³ and Lehmann et al.⁵⁴

Aggregating the cultural heritage sectors together is as problematic as combining the disciplines of the humanities into one would be.⁵⁵ This issue is amplified when considering metadata standards and how they are utilized across myriad research projects, institutions and by practitioners and researchers. Data in the cultural heritage domain necessarily must take many shapes as do the analogue 'originals.' This need is further exposed by the multitudinous metadata standards, as mapped by Riley,⁵⁶ at use in the cultural heritage domain. Moreover, continually changing digital processes and developing metadata formats lead to a fragility of digital resources over time, for example, early digitized content which is now getting past its prime.⁵⁷ Therefore, CHIs must continually review and have clear oversight of the changing technological landscape.⁵⁸ Digital resources, stored in the form of bits and bytes, may become obsolete if new technology implementations and systems no longer support

older data and metadata formats. This additional skill and knowledge burden placed upon CHIs could be adjudged to introduce fears that the adoption of new technologies and practices, if it was not governed by long-term strategy, could be in thrall to passing trends. Fragility of resources is of particular concern for digital-born materials and digital research data. As Vines et al.⁵⁹ discovered in one study on biological research data, 80 per cent of the scientific data collected by researchers in the early 1990s is no longer accessible. Whether this loss can be extrapolated to other domains at this level, or not, it can be considered a risk that was not taken into consideration at the time, and still exists if there is no strategic coordinated plan on the part of the cultural heritage sector to preserve digital-born materials.

The physicality of a CHI may instil a degree of trust in a researcher, which may diminish when research is conducted through the placelessness of the results of a search engine. For the researcher utilizing search engines for resource discovery, *caveat utilitor* must always be taken into account. The placelessness of such resultant spaces may present the user with knowledge which does not necessarily come with the trust one may expect with holdings of a well-established institution. New opportunities due to the spacelessness of the Web do not come without new risks such as trust in both directions. The digital future of CHIs raises questions about gatekeeping and control over more or less porous boundaries as the historical record becomes untethered from a physical location. Opening access to research data may result in a weak relationship between the institution and its users. Even though practitioners were concerned about the final interpretation of data, digital discoverability meant that their control over the representation of complexity was exercised further upstream of users' interaction with it than if they had engaged with the institution in person. This displacement of the relationship between user and cultural heritage expert in which a dialogue is nurtured (and questions are less likely to be ignored) poses risks to the best interests of both the researcher and the historical record.

This particular aspect of the KPLEX study found a desire for balance in which the historical record was rich with data that could be discovered and used without moral panics. Institutions were responding to new datascares of research by looking at their capacity to support changing research methods and investments were being made in technologies where it was hoped they would better serve practitioners adapting their practice. The changing nature of researchers' contact with institutions presents new challenges for ensuring cultural heritage knowledge is used to the advantage, and for the advancement, of the world's knowledge.

Notes

- 1 'Alphabet', *Alphabet*, <https://abc.xyz/>.
- 2 'Global 500', *Fortune*, <https://fortune.com/global500/2020/>.
- 3 This study further showed that less than 4 per cent of click-throughs happened on the second results page.
- 4 Petrescu, Philip., 'Google Organic Click-through Rates in 2014', *Moz* (2014), <https://moz.com/blog/google-organic-click-through-rates-in-2014>.
- 5 *Satisficing* is a portmanteau of satisfying and sufficing. It is a decision strategy of not necessarily selecting the best-possible option but the first reasonable option, since there is little risk in doing so in a Web environment. Krug, Steve, *Don't Make Me Think! A Common Sense Approach to Web Usability* (Circle.com Library, Indianapolis, 2000), 24.
- 6 Quantitative Social Sciences often utilizes well-known datasets, for example, for international comparisons, such as the Household Budget Survey (all EU states) and the Labour Force Survey (all EU states). However, qualitative Social Science may use data from many small sources including archived data.
- 7 Krug, Don't Make Me Think! A Common Sense Approach to Web Usability, 21–9.
- 8 Randall, Neil, *The Soul of the Internet* (Boston, 1997), 187–94.
- 9 Berners-Lee, Tim, and Robert Cailiau, 'WorldWideWeb: Proposal for a HyperText Project', <https://www.w3.org/Proposal.html>.
- 10 Nelson, T. H., 'Complex Information Processing: A File Structure for the Complex, the Changing and the Indeterminate', in *Proceedings of the 1965 20th National Conference*, ACM 65 (New York, 1965), 84–100, <https://doi.org/10.1145/800197.806036>.
- 11 Ibid.
- 12 Ibid.
- 13 A Web of linked data rather than a Web of linked pages. 'Semantic Web', *World Wide Web Consortium* (2015), <https://www.w3.org/standards/semanticweb/>
- 14 Berners-Lee, Tim, 'Semantic Web Road Map' (1998), <https://www.w3.org/DesignIssues/Semantic.html>.
- 15 Berners-Lee, Tim, 'Linked Data - Design Issues' (2006), <https://www.w3.org/DesignIssues/LinkedData.html>.
- 16 Kuhn, Tobias, Albert Meroño-Peñuela, Alexander Malic, Jorrit H. Poelen, Allen H. Hurlbert, Emilio Centeno Ortiz et al., 'Nanopublications: A Growing Resource of Provenance-centric Scientific Linked Data', *2018 IEEE 14th International Conference on E-Science* (2018) <http://arxiv.org/abs/1809.06532>, 83–92.
- 17 'Wikidata', https://www.wikidata.org/wiki/Wikidata:Main_Page.
- 18 'GeoNames', <https://www.geonames.org/>.
- 19 Maes, Pattie, 'Agents That Reduce Work and Information Overload', *Communications of the ACM*, 37/7 (1994), 30–40.

- 20 Jamsa, Kris A., and Ken Cope, *World Wide Web Directory* (1995).
- 21 Jansen, Jim, and Amanda Spink, 'How Are We Searching the World Wide Web? A Comparison of Nine Search Engine Transaction Logs', *Information Processing & Management*, 42 (2006), 248–63.
- 22 Wray, R., 'Google Celebrates Its 10th Birthday in September – or Does It?', *The Guardian* (2008), <http://www.theguardian.com/business/2008/sep/05/google.mediabusiness>.
- 23 Bryan, Kurt and Tanya Leise, 'The \$25,000,000,000 Eigenvector: The Linear Algebra behind Google', *Siam Review - SIAM REV*, 48 (2006), 569–81.
- 24 Chakrabarti, Soumen, Byron Dom and Piotr Indyk, 'Enhanced Hypertext Categorization Using Hyperlinks', *ACM SIGMOD Record*, 27/2 (1998), 307–18.
- 25 Brin, Sergey and Lawrence Page, 'The Anatomy of a Large-scale Hypertextual Web Search Engine', *Computer Networks and ISDN Systems*, Proceedings of the Seventh International World Wide Web Conference, 30/1 (1998), 107–77.
- 26 Bryan and Leise, 'The \$25,000,000,000 Eigenvector: The Linear Algebra behind Google'.
- 27 Kosala, Raymond and Hendrik Blockeel, 'Web Mining Research: A Survey', *ACM SIGKDD Explorations Newsletter*, 2 (2001), 1–15.
- 28 Singhal, A., 'Introducing the Knowledge Graph: Things, Not Strings', *Google* (2012), <https://blog.google/products/search/introducing-knowledge-graph-things-not/>.
- 29 'Refine Web Searches – Google Search Help', <https://support.google.com/websearch/answer/2466433>.
- 30 Edmond, J., M. Priddy, and N. Horsley, 'KPLEX Report on Historical Data as Sources' (2018), <https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:114127>, 14.
- 31 Browsing and the browse metaphor refer to the process of knowledge discovery whereby the users follow hypertext links (URI), to other web pages either within the same website or externally.
- 32 Cook, Terry, "'We Are What We Keep; We Keep What We Are': Archival Appraisal Past, Present and Future', *Journal of the Society of Archivists*, 32/2 (2011), 173–89.
- 33 Robertson, Hamish, 'Engines of Knowledge: The Museum and the Exhibit', *Discover Society* (2017), <https://discoversociety.org/2017/01/03/engines-of-knowledge-the-museum-and-the-exhibit/>.
- 34 Victoria and Albert Museum, 'Size of the V&A Collections' (2015), <https://archive.li/wip/QyjrP>.
- 35 A unique and constant identifier within the cultural heritage institute to allow a resource to be found. The name for this identifier varies depending upon the institutional culture and the form of the resource. For example, if it is a data file it would have a persistent identifier (PID), an archive folder or box may have a shelf marker, and a book may have a Dewey or Universal Decimal Classification.
- 36 In the cases of image archives, galleries, libraries and museums it is the object level description that takes precedence.

- 37 Edmond, Priddy and Horsley, 'KPLEX Report on Historical Data as Sources', 57.
- 38 Dodge, Martin and Rob Kitchin, *Mapping Cyberspace* (London, 2001), 13–17.
- 39 'Staatliche Museen Zu Berlin, Pergamonmuseum', <https://www.smb.museum/en/museums-institutions/pergamonmuseum/home/>.
- 40 'British Museum', <https://www.britishmuseum.org/>.
- 41 'About the Yad Vashem Archives', <https://www.yadvashem.org/archive/about.html>.
- 42 Dodge and Kitchin, *Mapping Cyberspace*, 14.
- 43 Ibid.
- 44 Ibid.
- 45 Universal Resource Indicator is sometimes known as an URL, Universal Resource Locator. It is the unique Web address for the resource (a web page): Berners-Lee, T., 'Semantic Web Road Map'.
- 46 The process and feeling of being instantaneously transported virtually to a new web or internet location without the context of the journey or relationship to the previous virtual space.
- 47 Kane, Arielle, 'Google the Gutenberg', *The Daily Pennsylvanian* (2009), https://www.thedp.com/article/2009/10/arielle_kane_google_the_gutenberg.
- 48 Even the digital photographing of objects may not capture the nuances of the object's medium, dimensions, construction methodology, etc.
- 49 Nauta, Garhard Jan, Wietske van den Heuvel and Stephanie Teunisse, 'Europeana DSI 2 – Access to Digital Resources of European Heritage' (2017), https://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/ENUMERATE/deliverables/DSI-2_Deliverable%20D4.4_Europeana_Report%20on%20ENUMERATE%20Core%20Survey%204.pdf.
- 50 Zuboff, Shoshana, *The Age of Surveillance Capitalism* (London: Profile Books, 2019), 67.
- 51 Ibid., 80.
- 52 Frankl, Michal, Michael Bryant, Jessica Green, Wolfgang Schellenbacher and Magdalena Sedlická, *Deliverable 12.2: Thematic Approach 1: Edition of Documents* (2018), <https://www.ehri-project.eu/sites/default/files/downloads/Deliverables/D12%202%20Thematic%20approach%201%20Edition%20of%20documents.pdf>.
- 53 Frankl, Michal, and Wolfgang Schellenbacher, *Deliverable 2.1: Terezin Research Guide* (2013), <https://www.ehri-project.eu/sites/default/files/downloads/Deliverables/Deliverable%202%201%20original.pdf>.
- 54 Lehmann, Jörg, Lorenza Tromboni and Andrei Zamoiski, 'D5.2 Archival Research Guides' (2015), http://www.cendari.eu/sites/default/files/CENDARI_D5.2%20Archival%20Research%20Guides_Final.pdf.
- 55 Borgman, Christine L., *Big Data, Little Data, No Data: Scholarship in the Networked World* (Cambridge, MA: MIT Press, 2015), 161.
- 56 Riley, Jenn, 'Seeing Standards: A Visualization of the Metadata Universe' (2010), <http://jennriley.com/metadatamap/>.

- 57 Hill, Mark J. and Simon Hengchen, 'Quantifying the Impact of Dirty OCR on Historical Text Analysis: Eighteenth Century Collections Online as a Case Study', *Digital Scholarship in the Humanities* (2019), <https://doi.org/10.1093/llc/fqz024>.
- 58 Also known as a technology watch when performed regularly.
- 59 Vines, Timothy H., Arianne Y. K. Albert, Rose L. Andrew, Florence Débarre, Dan G. Bock, Michelle T. Franklin et al., 'The Availability of Research Data Declines Rapidly with Article Age', *Current Biology*, 24/1 (2014), 94–7.