Emerald Insight

Journal of Documentation
Infrastructure as intermediation – from archives to research infrastructures
Sheila Anderson Tobias Blanke

Article information:

Users who downloaded this article also downloaded:

For Authors

If you would like to write for this, or any other Emerald publication, then please use our Emerald for Authors service information about how to choose which publication to write for and submission guidelines are available for all. Please visit www.emeraldinsight.com/authors for more information.

About Emerald www.emeraldinsight.com

Emerald is a global publisher linking research and practice to the benefit of society. The company manages a portfolio of more than 290 journals and over 2,350 books and book series volumes, as well as providing an extensive range of online products and additional customer resources and services.

Emerald is both COUNTER 4 and TRANSFER compliant. The organization is a partner of the Committee on Publication Ethics (COPE) and also works with Portico and the LOCKSS initiative for digital archive preservation.

Infrastructure as intermeditation – from archives to research infrastructures

Sheila Anderson and Tobias Blanke

Department of Digital Humanities, King's College London, London, UK

Abstract

Purpose – The purpose of this paper is to analyse the steps taken to produce new kinds of integrated documentation on the Holocaust in the European Holocaust Research Infrastructure project. The authors present the user investigation methodology as well as the novel data design to support this complex field.

Design/methodology/approach – The paper is based on the scholarly primitives framework. From here, it proceeds with two empirical studies of Holocaust archival research and the implementation steps taken. The paper employs key insights from large technology studies in how to organise such work. In particular, it uses the concepts of social-technical assemblages and intermediation.

Findings – The paper offers a number of findings. First from the empirical studies, it presents how Holocaust researchers and archivist perceive the way they currently do research in archives. It then presents how the intermediation and digital transformation of such research can be enabled without violating its foundations. The second major insight is the technical research into how to use graph databases to integrate heterogeneous research collections and the analysis opportunities behind.

Originality/value – The paper is based on existing work by the authors but takes this work forward into the world of real-life existing historical research on archives. It demonstrates how the theoretical foundations of primitives are fit for purpose. The paper presents a completely new approach on how to (re)organise archives as research infrastructures and offers a flexible way of implementing this. Next to these major insights, a range of new solutions are presented how to arrange the socio-technical assemblages of research infrastructures.

Keywords Archives, History, Interconnection, Knowledge management systems, Online databases

Paper type Research paper

1. Introduction

> But people don't do data well. Automated systems do, […]. We ought to remember a lot more from William Kent, about the ambiguities of concepts, but especially that bit about computers possessing incredibly little ordinary intelligence (Chris Rusbridge[1]).

The European Holocaust Research Infrastructure (EHRI; www.ehri-project.eu) aims to provide access to archives, to connect knowledge and to facilitate and enhance the process of research into the Holocaust in Europe and beyond. EHRI is developing an open, collaborative research environment, which provides integrated online access to dispersed (archival) resources relating to the Holocaust, across Europe, Israel and the USA, by linking archives, functionalities and people. EHRI has to work across this range of institutional settings, archives and other collection-holding institutions are embedded in. These can be large national archives such as the German Bundesarchiv but also a range of potentially less stable and more at risk institutions. Their funding is often limited and subject to political preferences, which means that at least some of their collections are hidden from view. Collections can also be hidden, as they are not found in the places one might expect them. Fundamental research collections on the Terezín concentration camp, for instance, can be discovered at the Dutch Institute for

War, Holocaust and Genocide Studies (NIOD; www.niod.knaw.nl), because one of the earliest historians to document the Terezín atrocities was friends with the founding director of NIOD.

Next to these kinds of archival dispersions that can be expected for many historical subjects, we can also find large-scale existing collection work on identifying and also preserving documentation specific to the Holocaust; mostly by Yad Vashem in Israel (www.yadvashem.org/yv/en/about/archive/) and the United States Holocaust Memorial Museum (USHMM; www.ushmm.org). Both have over the years brought together an impressive amount of documents and collections pertaining to the Holocaust. Mostly, this was done by photocopying the originals or even existing copies. Later, the photocopies were amended by digitizing the evidence. Washington and Jerusalem now hold large copy archives for research on the Holocaust.

Research into the Holocaust is as diverse as its archives and is often multi-disciplinary with researchers from across a range of disciplines including history, literary studies, sociology, etc. EHRI faces a significant challenge then in form of the sheer amount of information, the range of cataloguing approaches and technologies, the disparate nature of the material, which includes documents, letters, photos, films and art, and the range and variety of the research topics and questions addressed by Holocaust scholarship.

This paper analyses the systematic investigation by EHRI into current practices and processes of Holocaust research and the subsequent work to translate these into a flexible technical infrastructure. We proceed by first theoretically placing our work in the context of research on infrastructures and offer insights into the framework we have developed to link research practices with infrastructure services through scholarly primitives. We then depict our methodology used to collect data on research practices; followed by a discussion of our analysis approach. Finally, we present how we employ a specific set of database technologies and graph modelling to develop a set of services that constitute together our integrated research infrastructure.

2. Infrastructure – continuities and intermediation
Holocaust archives are involved in a process of deep transition from established analogue practices of documentation towards the possibilities of digital representation and mediation. Hayles (2007) offers the concept of intermediation as a possible theoretical framework through which we might test what happens when we move from one medium to another – in this case from the analogue, physical spaces of archives to the digital, distributed spaces of research infrastructures. Hayles (2007) argues that "knowledge is carried forward into the new medium, typically by trying to replicate the earlier medium's effects within the new medium's specificities".

Infrastructures are often the means to carry knowledge forward into a new medium. They are socio-technical assemblages to develop a new medium and enable intermediation. "[I]nfrastructures must be understood in their entirety, as hybrids that join and rely on elements too often separated under the (bogus) headings of "technical" and "social", etc. Transformative infrastructures cannot be merely technical; they must engage fundamental changes in our social institutions, practices, norms and beliefs as well" (Edwards *et al.*, 2012, p. 12). Research into infrastructures is then research into the assemblages of the technical and social "headings" they consist of.

Hayles's concept of intermediation is a good starting point to understand the particular challenges for infrastructures that invoke a technology change in the social fabric of a society or a research discipline. We need to avoid the mistake of over-stating the digital revolution in research and re-think infrastructures not only from the overall revolution

they entail and the grand vision of global digital spaces but from the current practices and how these change with digital research infrastructures. Only like this, infrastructures can provide their primary function of continuity. In the process of intermediation, infrastructures suggest stability: "[I]nfrastructures are largely responsible for the sense of stability of life in the developed world, the feeling that things work, and will go on working, without the need for thought or action on the part of users beyond paying the monthly bills" (Edwards, 2003, p. 188).

Blanke *et al.* (2013) has elsewhere argued that, in order to provide stability, research infrastructures should be considered as incremental, as arising from the already established procedures and practices of both archives and scholars, whilst at the same time enhancing these practices through an interaction with the capabilities of the computer and computational processes. Whilst the goal for humanities research infrastructures remains to integrate disparate and dispersed sources, it must do so in the context of the process of knowledge-making as driven by and for the purposes of research and with the computer as a crucial component of the infrastructure. Only in this incremental way, Weinberger's (2011) prediction becomes reality that for research infrastructures, "the smartest person in the room is the room itself: the network that joins the people and ideas in the room, and connects to those outside of it […]" (p. VIII). The question we would like to answer in this paper is how to design a room that is smarter than the people it contains – especially if the room contains many smart researchers. The room needs to allow for Hayles's intermediation to take place to support the stability that Edwards demands for infrastructures.

The concept of scholarly primitives (Unsworth, 2000), fundamental research processes common across disciplines, has proved popular in providing a framework with which to translate research practices into a set of activities and underlying primitives that can be used to inform infrastructure design and functionality. In 2002 Unsworth (2000) suggested that the term primitives could be used "[…] to refer to some basic functions common to scholarly activity across disciplines, over time, and independent of theoretical orientation". Since then, perhaps the most widely quoted synthesis and discussion of primitives has been provided by Palmer *et al.* (2009). They identified five categories of scholarly activity: searching, collecting, reading, writing and collaborating; together with a further set of cross-cutting primitives that took place across all the categories. Blanke *et al.* (2013) were the first ones to use the scholarly primitives provided by Palmer as a conceptual framework for developing (institutional) infrastructures for humanities research work. They concluded that to ensure relevance for infrastructure work a revised activity category set was necessary that included: discovering, collecting, comparing, delivering and collaborating, together with the finer grained primitives that sit underneath these categories, chaining and browsing under discovering, for example.

For this work, scholarly primitives help design the assemblages of the technical and social that research infrastructures are and support our studies to investigate the relationship of users and infrastructures. Pollock and Williams (2010) conclude based on ethnographic studies of existing infrastructures that their decisions need to be formed "by the specific research concerns and issues under examination". We are interested in primitives as they help us understand the role of these research concerns and issues, while being generic enough to accommodate different needs. They summarise research activities within the context of the life cycle of digital research objects within them and within the technical and social assemblages they are embedded in. As EHRI is a new infrastructure, we are, however, not able to observe existing behaviour. We cannot engage in ethnographic studies, as Pollock and

Williams did. We thus rely on our own past work and the theoretical derivation of primitives by Unsworth and others to guide our investigative studies in how to best realise a research infrastructure for Holocaust studies. Primitives provide a good way to avoid developing an infrastructure for the sake of an infrastructure, which Rockwell (2010) sees as a danger in such projects because primitives link "researchers" processes closer to the development of services and techniques (Schreibman *et al.*, 2013).

3. EHRI primitives and methodology

"As science has gotten too big to know, we've adopted different ideas about what it means to know at all" (Weinberger, 2011).

The theory of scholarly primitives grounds our research and investigation work. It provides the guideline to group and code our interview data and analyse our surveys. In our work for EHRI we found five overall categories that are constituted of 15 individual primitives as listed in Table I. These categories were partly taken from existing work and partly adapted.

The first group of primitives we called Searching, which is fundamental to archival research work and to finding secondary published sources, which inform and shape the construction of new knowledge. However, Duff *et al.* (2012) suggest that instead of assigning this behaviour as "searching" our understanding would be enhanced were we to regard it as an essential part of researchers' meaning-making, and as a component of the interpretive process rather than "seeking" activity.

Unlike Blanke and Hedges (2013), we included "reading" (in the wider sense) as a core activity, as our focus is the integration of archives and research on the Holocaust. We reinstated reading as the process of reading, assessing and re-reading proved an essential component of archival research. Here, "reading" is interpretive and iterative rather than a linear process where records are searched for, read and then interpreted. Cross-cutting primitives were also found to form essential components of research work. This fits with the argument in (Duff *et al.*, 2012) that the activities undertaken as part of archival research work are an essential part of the meaning-making process.

In this paper, we concentrate first on how scholarly primitives help us identify the patterns in the interviews we can use to develop services. Scholarly primitives are useful to understand the key concepts of research practices in the interviews that drive

Table I.
Palmer's scholarly activities and primitives extended for EHRI purposes

Searching (investigating)	*Collecting*
Direct searching	Gathering
Chaining	Organising
Browsing	
Probing	
Accessing	
Reading (interpretation)	*Collaborating*
Scanning	Networking
Assessing	Consulting
Re-reading	Sharing
Cross-cutting primitives	
Note-taking	
Translating	

archival research. While in our interviews and other studies we should have an open mind about scholarly work processes, this does not mean we need to keep the "head empty" as Dey (2007) emphasises for qualitative research. We can use our prior knowledge of scholarly primitives to guide our investigation and structure our understanding.

Our studies to identify and understand research practices in the field of Holocaust research began with identifying and managing the Holocaust research landscape through a literature review and annotated bibliography. Based on the results of this survey we identified a range of researchers, archivists and other people to conduct our investigations with. The survey questionnaire was designed to complement the interviews following Glaser's famous idea that "all is data" in qualitative research (Glaser, 2002). The interviews were semi-structured to allow for the detailed collection of the interviewee's views and actions and to pursue further discussion of interesting lines of enquiry.

3.1 Study 1 –survey
Our first study was an online survey for Holocaust researchers undertaken over an eight-month period with a total of 277 responses. The responses were collected between October 2011 and May 2012. Most respondents lived in Europe, while 14 per cent lived in the USA and 6 per cent in Israel. The questionnaire comprised 17 questions, which were expressed in either binary nominal (yes/no) or five-scale ordinal (Likert) scale and covered the following themes: details of the respondent including demographic information, research area and discipline; resources used; activities and methods; procedures, beliefs and attitudes; tools and services used; goal and motives.

3.2 Study 2 –interviews
Next to the survey, we conducted eight semi-structured interviews with archivists and librarians working in a variety of archives as well as 15 semi-structured interviews with researchers. Potential interviewees for the researcher interviews were obtained from the bibliographic analysis but also from recommendations by the EHRI partners. Based on this theoretical sampling of the body of those involved in research on the Holocaust, a total of 24 researchers were approached and 15 provided extended interviews. Researchers were selected to cover as wide a range of disciplines as possible, and to represent different career stages, ranging from doctoral candidates to full professors. Interviews lasted between 45 minutes to two hours and were digitally recorded, transcribed, encoded and analysed using Nvivo. The scholarly primitives framework was used to structure the content analysis.

4. Study results
4.1 Searching and investigating
As described, archival research is not a linear process but is recursive and interpretive. The ability to find and assess sources is highly variable, some are easy to locate whilst others require a long and detailed investigative process. Researchers move quickly beyond simple "searching" techniques and instead investigate and track down sources that may be of interest, following hunches and being open to serendipitous discovery. "Searching" is, as presented earlier, part of the meaning-making process with interpretation and assessment of the sources included from the beginning.

Once our studies looked into the specifics of searching for research in archives, search strategies of researchers tend to coalesce around faceted searching, with names, dates, places, events and subject being the most popular:

> Let's say I've got 3 or 4 important words like Germans, Jews, Vichy police, and, after that, depending on the topic I am using all the words. So currently I am working on the alienation, so I am using this word, or selling of Jewish goods, or Jewish assets[…] (Interview with Holocaust researcher).

The key concept expressed here is "topic-based" searching where topics describe specific research interests. This result from our interviews is consistent with our prior work in (Blanke *et al.*, 2011) that investigated faceting searching and its relationship to research processes. The prior work has demonstrated that faceted searching and browsing and the concentration of concepts that can be flexibly combined is a good way to allow for new research questions to emerge.

A key difference of the research process compared with other more standard searching and browsing activities is that questions and information needs are (re-)formulated while investigating the data based on the subjects found in the material:

> We [archivists] think of the [archive] as different collections; they [researchers] think of the [archive] as different subjects (Interview with archivist).

As a result of this research need a range of other databases next to traditional finding aids can now be commonly found in most archives, though finding aids continue to be one of the most popular retrieval methods (Anderson, 2004). Daniels and Yakel (2010) report on retrieval experiments with users of archival information systems that demonstrate the range of search activities in contemporary archives, while Anderson and Allen (2009) see an archival commons developing using modern networking technologies. Ross (2012) finally discusses the wider use of information technologies in archives to preserve and organise its records.

Our interviews with archivists illustrated that many less experienced researchers find understanding the way, in which an archive is arranged and how the finding aids are ordered, difficult to comprehend (Speck and Links, 2013). This resonates with the findings of Yakel (2004), who investigated the research use of the typical archival hierarchy of repositories, collections, files, etc. She found that researchers struggle with archival terminologies. Sometimes search functionalities are simply completely missing and definitely not working across the range of archives with their varying content and formats:

> Like I said, there is no basic search possibilities through all the archives. That's the most simple thing [we could do to improve things] I think (Interview with Holocaust researcher).

Despite these drawbacks, the majority of researchers indicated that they were reasonably well acquainted with archival search and in general had no grave concerns regarding their navigation or orientation within an archival environment. Some expressed at least some difficulty in using an archive's catalogue:

> I think you have to gain the trust of the archivist in some ways and in a lot of archives the cataloguing is very poor so there's not any really decent catalogue so you just have to really trust the archivist to give you the things that you need (Interview with Holocaust researcher).

The keyword here is trust in archivists but also other researchers using the archives. More than 50 per cent of the researchers expressed in our studies that they would put time and effort into cultivating a relationship with an archivist. Successful establishment of these

relationships was considered especially useful in tracking down less well-catalogued materials or to point to materials of interest held in other archives. Archivists have therefore got an essential function in researching new findings and materials.

Trust into archives and their holdings (Ketelaar, 1997) is widely discussed in the literature and especially the relationship between historians and archivists (Duff and Johnson, 2002; Case, 1991). Johnson and Duff (2005) pointedly call this relationship the "social capital" a historian has to develop to succeed in her research. The authors recommend that in order to build trust, "with the increased use of the World Wide Web, archives will have to create systems that will duplicate the archivist's contextual knowledge of collections that is so valued by historians" (Johnson and Duff, 2005, p. 129).

Similarly, Anderson (2004) could identify how researchers ignored more advanced digital retrieval techniques and relied on their trusted sources to locate material in archives. Over 30 per cent of his interviewed historians even thought that online retrieval would be the "least effective method" of finding material in archives because they did not trust the accuracy of the search engines. In order to counter such concerns, Anderson (2004) therefore demands future archival information systems that are more than search engines and are also "expert systems" and "intelligent agent[s] able to conceptualize, mediate, and tailor the information provided" (p. 114). Anderson quotes Amazon's digital library as a contemporary example for such an intelligent digital agent. These kinds of third generation digital library systems have also served EHRI as inspiration for some of our implemented features for working with and building collections from digital archives.

4.2 Collecting

Once the archival material is found, it needs to be collected and then prepared for analysis. This can mean to re-create indexes for archives in order to better match research needs. New research collections develop. Our survey showed strong divergence of opinion concerning the usefulness of current archival information retrieval systems for this collection-development activity.

The investigative process on the Holocaust necessarily extends to working across multiple archives and using multiple sources in the process of their research. We have described in (Blanke and Kristel, 2013) how the transnational aspect of the Holocaust as well as the dispersal of its research material is in fact one of the main motivations behind EHRI. This focus on collecting material from many different archives is a common feature for most archival research but the Holocaust is also a specific case with its wide range of countries involved as well as bodies of the state:

> I'm using a great variety of sources including archival documents, so those are documents of the public administration and law enforcement agencies and so on, and also trial material, so documents of prosecutors and trials post-war mainly (Interview with Holocaust researcher).

The key concepts in this statement refer to the great variety and different types of research sources in Holocaust research. This confirms the necessity for the overarching goal of EHRI to integrate archives and sources by connecting concepts and knowledge elements across archives and collections where possible and at the deepest level of granularity.

As discussed above, EHRI concentrates on the "hidden collections" (Blanke and Kristel, 2013), which were cited as significant by almost all researchers and archivists we surveyed. They can be hidden from research and public view for many reasons. They might not be catalogued and discovered yet, they are subject to access

1189

restrictions or they are in a language neither researcher nor archivist understands. Interviewees from four organisations pointed out this issue to us, which seems common with the highly dispersed and distributed Holocaust collections. There are even institutions that offer partial translation services to mitigate this problem. Furthermore, over 90 per cent of the organisations interviewed had collections with access restrictions, and more than half reported on large amounts of collections that were not catalogued at all.

We noted elsewhere (Blanke and Kristel, 2013) that especially these hidden collections interest researchers as here they might collect from here the new research material they can publish on. This puts researchers again in a strong dependency with the archivists, as they are often the only ones who can point them to the hidden collections. On the other hand, once the researchers have created their own indexes of these hidden collections, they are not shared with the archives for enhancing their databases. Researchers' collections can be lost as they are often not formally organised and registered. EHRI has organised several workshops on the early post-war documentation of the Holocaust that demonstrated the need to consider the time collections were created and also how they developed from that point on. The early documentation efforts were often done by survivors and are today themselves subject of research.

Contemporary research documentation uses increasingly personal databases, where Holocaust scholars create their own research-oriented documentation specific to their subject of interests. The following quote is typical for contemporary researchers, who will over time accumulate a number of databases for their own collections. All of these become part of the wider body of research material on the Holocaust, a modern form of grey literature:

> [T]here is one database per person, like a personal folder for each person, […], another database is the aliases, because this particular archive was only about aliases and you get crazy with it, and another database with the concepts, the institutions, because there are so many agencies that I couldn't remember which is which. But I also keep records in Word were I prepare the index (Interview with Holocaust researcher).

Key concepts that help us understand researchers' needs in this statement are personal folders and research-led indexes. Though the indexes are often not shared with the archives, the situation is better when it comes towards sharing with fellow researchers.

4.3 Collaborating and sharing

Our studies show that historians working in archives consider sharing of findings an essential part of their research work. Collaboration is necessary when faced with kilometres of archival material:

> […] if I find these other things in the archive that I know will interest a colleague I will send it to them and they will send me what interests me (Interview with Holocaust researcher).

Sharing of findings then leads to a collaborative research process and interaction among researchers at every stage of the analysis:

> […] From the beginning to the end I speak with people about the resources I can use, about the analysis, and even after that when I write […] (Interview with Holocaust researcher).

The key concepts in these statements are collaborative analysis and sharing of findings. In this sharing process, the research community includes archivists. One of

our interviewees illustrated how she readily works with an archivist to share information:

> […] he and I run the book collection and he is pretty much engaged in the knowledge sharing, what kind of tools you can use on the Internet for sharing knowledge and information, like del.icio.us and so on (Interview with Holocaust researcher).

We conclude from this that EHRI has to provide not only an environment through which researchers can find, retrieve and use EHRI content, but also a social space in which EHRI researchers can share and exchange knowledge, expertise and information. In short, EHRI should aim to operate as a digital research ecosystem (Anderson and Blanke, 2012).

The relationship between archivist and historian is, however, not always straightforward. Discussions of the "archivist-historian" (Bolotenko, 1983, p. 20) go back to the early days of the internet and computing or the "electronic eighties" (Taylor, 1984, p. 30). In the discussions, the main concern was whether the emerging information technology required the archivist profession to change fundamentally, depart from its traditional link with the discipline of history and develop into "overarching information generalists with an archival emphasis" (Taylor, 1984, p. 32), who would only prepare "intractable historical records" for scholarly research. In our studies, we found both archivists, which consider themselves information generalists, as well as archivists, which think of themselves as historical experts on their sources and can provide researchers with contextual knowledge. Historians in our interviews suggested that they preferred the latter kind of archivists compared to those only concentrating on information management:

> I think that […] archivists don't always want you to look at the material (Interview with Holocaust researcher).

While the historian often has no choice, the preference is clearly the archivist who is an "archivist-historian" and is "a representative of the world of research in the world of administration"; recognising that archives share the care for sources with rational modern history. "[A]rchivists must transcend mere information, […], if they wish to search for, and lead others to seek, "knowledge" and meaning among the records in their care" (Cook, 1984, p. 49).

Our research has confirmed that in the "electronic age" scholars using archives expect more contextual knowledge and consider the information management to be less important. We believe this is linked to the fact that information retrieval systems in the twenty-first century rely on extended contextual knowledge in order to discriminate relevant from non-relevant knowledge. For an archive with a specific organisation acquisition and organisation mandate, on the other hand, the information professional might be more important, wherein lies the continued tension between archives and researchers long after the "electronic eighties".

4.4 Note-taking/annotations

Our studies demonstrate that the vast majority of both archivists and researchers believe that enhancing, adding and annotating resources is important (with the appropriate control mechanisms). One of the archivists interviewed even suggested that annotators should be registered through Facebook so they would be traceable. Similarly, researchers wished to control what they shared and to be able to limit who they shared with. Roughly two-thirds of our survey respondents expressed a wish to

find out about other researchers current research work and almost as many stated they would be prepared to share resources and information on their own work with others.

Annotations can grow and develop into independent information resources that are often kept private:

> I have many electronic databases. But I also keep cards for bibliography and thematic cards, that I keep at home and haven't written them in my computer (Interview with Holocaust researcher).

This key role of researchers' own indexes and annotations is recognised by archivists and, indeed, we found cases where the annotations and indexes are used to enhance the archivists' knowledge of the archives in their care:

> Annotations [are] always important, and good annotations, because sometimes people are coming here and say, you've got something in your archive and what does it mean? And I also don't know what does it mean (Interview with archivist).

But there seems to remain a barrier of integration of the annotations and indexes into the formal knowledge of archives. Rarely, is this knowledge formalized and stays instead on the computer and notebooks of the researcher and is only partially published in articles and books as the outputs of research.

The introduction of computers and computational processes into an archival infrastructure offers the potential to leverage access to the outputs of the process of research. During their work in archives researchers build up a body of knowledge on the archives, the archival records and the finding aids. Once found and retrieved, researchers create associations between those sources, using annotations, mark-up and notes. In the process, they construct a significant personal collection of archival materials carefully referenced back to the original archive. Alongside the explicit knowledge incorporated into these personal collections, researchers will have gained a tacit knowledge about the archives, archival research and how to translate research questions into formal searches that yield results in the form of records and documents to answer their research questions. This can become a complicated and sophisticated research workflow:

> I have two electronic bibliographic databases in Access, one for modern history and one for the older one, author, title, place of publication, year, publisher, […]. I have made a full indexing. And then I have 3 databases for my work, about what I do, for the processing of a file that was about the agents this service had, per research topic. This database contains 2-3 other databases (Interview with Holocaust researcher).

These research databases, which realise the key idea of self-indexing, are the direct result of reading and scanning activities.

4.5 Reading/scanning

At each stage of the research process archival traces are discovered and discarded, created and noted, until such time as the work coalesces into a wider understanding of the topic under consideration:

> I think a better term than "searchprocess" for archive research is "investigative process", because the researcher really is acting like a detective or private investigator. Instead of searching-and-finding, the process involves a lot of back-and-forth work (Interview with Literature Researcher).

This interview we conducted with a scholar we met in the context of a Holocaust archive but who is not directly involved in Holocaust research. We wanted to determine the interdisciplinary nature of the field and gain an outside perspective on potentially new approaches. Going through archival material is for the researcher like detective work and involves the close reading of the documentation involved. This reading of the archival data is characterised by two typical stages, also found in other research processes. First, "initial scouting" will deliver an overview of everything that is there. Here, the researcher wants to get everything and is very focused on the recall of her investigative work. This stage is followed quickly by deeper reading activities, which are characterised by a "back-and-forth" reading activity. Together these two stages describe much of the research process, before material is added to notes and then finally included in formal publications. Yakel (2004) presents a similar process but concentrates less on the individual reading and scanning steps.

Archival research guides provide a narrative integration of collections. EHRI created a range of research guides focusing on the Terezín/Theresienstadt ghetto and the Jewish communities in East-Central Europe (Speck *et al.*, 2014). Researchers reported positive experience with research guides beyond the context of their immediate work on Holocaust sources:

> Particularly useful was Colin Thom's Researching London's Houses: An Archives Guide, which gave me precise instructions on what resources the London Metropolitan Archives held for research into buildings, occupants and street plans (Interview with researcher).

It seems that researchers deal particularly well with this kind of more narrative integration of archival sources, which is seen to give precise instructions for their need.

Figure 1 summarises the EHRI primitives and the key concepts in the interviews that point to them. It also links primitives to key concepts, as mentioned in the studies. This does not mean that there are not also other links but this is how we met them in
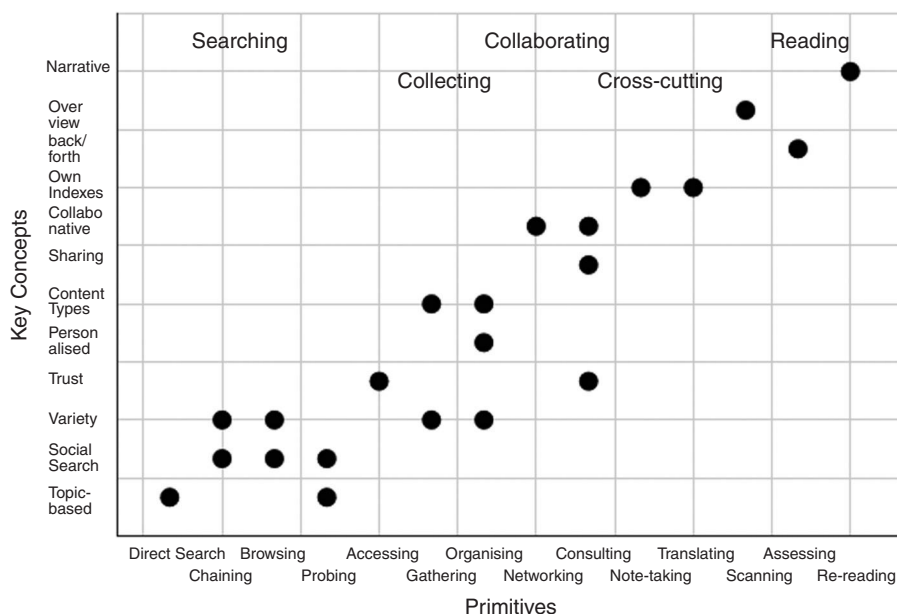

Figure 1.
EHRI primitives and key concepts

the interviews. As a result of our studies EHRI does not only integrate archives and sources and provides cross-searching of archival descriptions, but enriches these with the addition of researcher-generated annotations and classifications. It encourages and facilitates researchers to share their sources and links alongside the research questions and research topics for which they were selected and incorporate these into the EHRI finding aids. This enables in turn a new kind of reading of archival material, for which we needed to develop a new data infrastructure.

5. From primitives to data infrastructures – research infrastructure as an intermediary

For EHRI, the technical design challenge was to innovate a dynamic, research-driven environment, where new material is permanently discovered, added and analysed. We needed to rethink some of our assumptions that stemmed from earlier work on relatively stable cultural heritage collections (Blanke *et al.*, 2013) and develop an environment that is flexible enough to allow for the integration of heterogeneous material and that is social enough to allow researchers to discover and analyse their material and make new connections. "In building cyber infrastructure, the key question is not whether the problem is 'social' or 'technical' […]. The question is whether we choose for any given problem a social or technical solution – or some combination. The distribution of solutions is the object of study" (Dutton and Jeffreys, 2010, p. 43). We would add to this from our perspective that it is the assemblages of "solutions", which makes research infrastructures possible and a reality. Scholarly primitives help define this assemblage as well as the "solutions", because they bridge the technical and the social.

We cannot describe detailed components of this infrastructure here, as we have done so elsewhere (Blanke *et al.*, 2013). In the paper, we discuss how we can avoid setting up an infrastructure that makes a priori decisions on how the data for researchers is supposed to look like. Researchers want to make their own decisions about their material and its relevance themselves, often create their own views on the information they get from the archivist and simply want to be as flexible with the data as they can.

In this paper, we would like to rather analyse how the primitives have made our work on the data infrastructure possible and helped us translate the work of researchers and archivists with material in traditional archives into a data-reading digital infrastructure. Using the primitives we could address the intermediation concerns of Hayles as well as the Edwards's demand of stability for infrastructures. We need a data infrastructure that can change with the new questions that digital research wants to enable without losing data or making complicated changes to its formats and schemas. As Gitelman (2013) states, "raw data" is an oxymoron, and it needs to be argued to "provide a rhetorical basis". To provide this rhetorical basis and escape the earlier cited expectation by Rusbridge that "humans don't do data well", data needs to be read and it can only be read if it is brought together. We needed to find a way to keep the researcher close to her data and decided to use NoSQL technologies (Sadalage and Fowler, 2012).

In particular, we investigated graph databases for their ability to integrate dynamic, heterogeneous and evolving data sets:

A unifying database built using a Graph DBMS has no need of a mediated schema or any other consensus structures. The anything-can-connect-to-anything nature of graph-based storage means that every table and column name from every schema in every data silo can be retained in a graph representation of enterprise data[2].

While we believe that it is too strong that graph databases have no need for schemas, it is correct that they are much more flexible than other data infrastructures when it comes to bringing together digital material. They can really connect (almost) anything to anything. Graph databases also work well with the typically hierarchical data in archives, they easily integrate social aspects into the data work and consider relationships as first-class citizens of the data model, which is especially key in historical research, as we describe in Blanke *et al.* (2010).

Next we demonstrate, how the archival research primitives map to the particular graph database we use. Neo4j[3] is an open source graph database that was first released in 2010, at about the time we started EHRI, and has since seen a wide range of applications in research and industry.

### 5.1 Searching and investigating

Neo4j adheres to the property-graph model, which we can exploit to enhance the researchers' data searching and investigating experience, since the graph database allows us to model our data in a manner that closely conforms to the actual relationships between entities. The graph structure greatly simplifies dealing with hierarchical structures, so typical to archives, compared to relational databases and offers new possibilities of exploiting the graph structures for research (Blanke and Kristel, 2013).

In order to support efficient chaining and browsing of research material from the great variety and different types of documentary evidence for the Holocaust, we have serialised certain key relationships; for instance, an archival description representing a particular series, and its parent fonds. Because we can infer from the organisation of archives, that the number of ancestors, a given hierarchically positioned item has, will not be unbounded, we know it is safe to automatically serialise an item's parent, and its parent's parent, up until the top level. This way, we maintain the archival context of Holocaust research material while chaining a documentary unit to its context. Graph databases such as Neo4j are designed to make these kinds of traversals easy, and "looping" through relationships with an unknown number of total steps is a fundamental feature of graph query languages (Sadalage and Fowler, 2012). Browsing through material in Neo4j thus implies walking its graph and reading its data.

The next steps in order to enhance this chaining and browsing experience include for us more advanced visualisations of relationships between documentary evidence that aid the investigation. We have just released, for instance, a visualisation of all provenance trails for archival documentary units (see Figure 2), because the meaning of research material is closely related to how, where and when it was created (Bechhofer *et al.*, 2013). A complete network analysis will be part of future work. In (Blanke *et al.*, 2013), we describe an initial experiment with a "shortest path" algorithm, that determined if two users had interacted with any of the same archival items, by whitelisting a set of relationships for the algorithm to traverse until it either finds one "user" node from another, or runs out of relationships.

Topic-based retrieval and probing are key concerns of our work. One of the advantages of graph databases is their ability to integrate different data types and models. Any data can be stored as edges and nodes between them. In our case, this implies that archival content is part of the same data infrastructure as archival metadata and as support structures such as thesauri. EHRI (Blanke and Kristel, 2013) has developed an integrated thesaurus to bring together terms from its archives.

Figure 2.
EHRI provenance
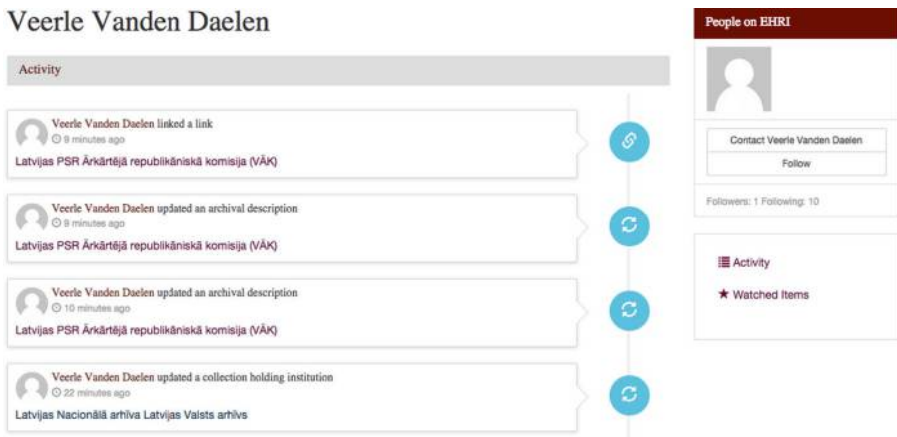
EHRI also provides computational access to topics in the collections. Next to the graph database we offer access to the collections through the Apache Solr Search engine. The Solr index goes beyond the native graph database index by providing the textual and linguistic properties of archival description data. Based on our experimentations with information extraction from OCR'ed collections of Holocaust research (Rodriquez *et al.*, 2012), we believe that there are many opportunities to enhance research access to the collections using topic-based query expansion and various other advanced techniques such as topic-modelling of descriptions.

5.2 Collecting

The EHRI data infrastructure gathers and organises manually created collections by researchers involved with the project as much as automatically ingested collections from archives, which provide a corresponding API. It is a distinct feature of our research infrastructure compared to a more traditional digital library that we do not aim to collect as many collections as possible in EHRI but those of a higher research value. This means that we characterise the collections in EHRI according to the detail of their descriptions and rank higher those collections that have been created by our own researchers and researchers involved in related projects. The research value as measured by the provenance of the documentary evidence as well as the detail of the description involved boosts the rankings.

The whole EHRI graph makes the collecting and integrating of archival evidence as easy as creating new nodes and relationships between them. As long as data can be mapped onto the property-graph model, any new data model can be added to the existing data. In the context of semantic web applications, interoperability is achieved by mapping all data items as triples of subject, predicate and object. Similarly, in graph databases a single data model dominates, but with the distinction that relationships are modelled explicitly and can have properties themselves (Sadalage and Fowler, 2012). Thesaurus terms, where relationships represent their typical narrower/broader view between concepts, can be new nodes as much as annotations or other user additions.

New data sets and different types of content can thus be added to this environment by mapping their data model onto nodes and edges between them. This means that the

environment scales towards information that is not uniform. Researchers can gather sources and combine them with concepts, using their own mark-up and notes to provide a personalised view on the collections. For a researcher, this kind of environment that can be dynamically expanded and where adding new collections and annotations is easy, supports optimally the close work with complex and wide-ranging material.

However, if new data items can be added easily, performance is critical. While Neo4j's design scales well with greatly increased quantities of data against user traversals (Maharajan, 2012), we still had to include further improvements for the performance based on assumptions how data will be added in archival research environments. Neo4j performs well with an increased complexity of the collections involved, as relationships are not resolved at query time like in relational databases but stored in the database. Entry points to the graph are located with specific indexes, and the speed of traversal from those entry points of a typical then localised (or "egocentric") query remains constant with regard to the total number of nodes in the database. As we also preserve the hierarchical relationships typical in archives, we further improve performance (Blanke *et al.*, 2013).

5.3 Collaborating and sharing

Nielsen (2012) describes how networked research requires collaboration. This is the case, not just because complex material requires consulting, as "even Albert Einstein needed help occasionally"[4] but particularly in sensitive fields where information needs to be permanently reviewed for disinformation as in Holocaust studies. Collaboration is critical to peer review in order to build a trusted network.

EHRI implements standard features, one has come to expect from peer collaboration applications. There is a forum; a series of profile pages and one can even follow the activities of fellow users while they are working through the EHRI data sets. The networking with other users is mainly determined by working together on similar collections and research objectives. Researchers can find out who is interested in similar material and share together "virtual" data collections. All of these functionalities are well supported by graph structures, where user collaboration as well as virtual collections constitute simply another form of nodes and relationships between them.

Trust in the material but also in the activities of fellow EHRI users is provided on the one hand side by a rich profile page for each user but mainly through transparent provenance trails of all work done on the EHRI graph. The source repository is recorded and matched against a list of authority archives, as is every action that is undertaken subsequently within the EHRI environment. Figure 2 demonstrates how the latest activity is recorded and visually discriminated.

But social collaboration between researchers in EHRI is not limited to the EHRI environment directly. The graph structure facilitates the easy integration of social data from researchers' LinkedIn or Facebook profiles; especially since Facebook has begun promoting its own graph search methodology. In order to find best possible ways to connect to the knowledge with fellow researchers and archivists, we experimented with the approach outlined in http://blog.Neo4j.org/2013/06/, providing links between sample researchers' and archivists' Facebook profiles and their work on sources in EHRI. For the experiment we asked two researchers to download their Facebook data using the "Give me my Data" tool and imported it into Neo4j. We then let them discover some of their Facebook friends and link their work to existing sources in EHRI. In a real research scenario, we would assume that these connections would have been there through the past work of the other researchers and archivists connected through the

Facebook friend graph. This way we enable an exploration of the sources that could follow paths like "Give me sources my colleagues (from Facebook) looked at". It would also allow to find all additions researchers have done to the EHRI graph.

5.4 Cross-cutting work

In Blanke *et al.* (2013), we describe how researchers and archivists can add annotations, notes and relationships in the EHRI environment. These can be shared or kept private within a community of researchers to develop virtual collections and views on the EHRI collections. In this section, we would like to concentrate on the graph implementation decisions we had to take in order to integrate user additions to the EHRI graph efficiently.

To capture user events in the EHRI graph such as new annotations and notes, we added a data structure that provides efficient recording of events by means of a global linked list of event nodes. This list is temporarily ordered with the latest event always the head of the list. This global event list is linked to from each node in the EHRI graph through stub notes that point to the relevant event in the global list. A linked list for global events can be much more easily implemented with graph databases than with relational ones where linked lists have to be realised through explicit pointers to header and next items. Querying and updating items in a linked list in relational databases is complex and potentially time-consuming.

For provenance trails of these annotations, we are interested in the latest events that occurred. We find all the latest events on the top of our global linked list of events and do not need to compare timestamps. Finally, the global event list presents three different views on the state of the EHRI archive (Blanke *et al.*, 2013). The global view provides an overview of global updates to the EHRI graph, the item-level view presents events such as annotations against a single item, while finally we have a view, which represents all the actions of a particular user in order to show what a user has done lately. All of these views facilitate the scholarly activities of scanning of relevant updates in the EHRI archive.

5.5 Reading/scanning

We have already discussed, how graph databases put relationships between information to the foreground. With their emphasis on relationships, graph databases are particularly well suited for historical research in particular and humanities research in general, as they translate typical scanning and reading activities into a data structure; the ability to flip back and forth between documentary evidence and the desire to get a quick overview. All the technologies described above support this reading of data that transforms it from its raw state into meaningful relations. But again efficient processing will become an issue with larger numbers of collections. In order to provide efficient scanning, we therefore reified a relationship into a node, if the relationship had a lot of additional metadata. Additional research is, however, needed into how this process can be formalised.

As we have already presented how graph databases support efficient reading and scanning access, here we would rather like to concentrate on another specific implementation and view on the EHRI graph that we provide. Research guides focus on the narrative integration of material for a particular topic such as the Terezín concentration camp. In Speck *et al.* (2014), the complex integration challenges for the research guides are analysed from heterogeneous metadata to the different systems used by the partner institutions. The guides integrate Holocaust material not just along

one axis from collections to files but by developing the story around them and with multiple entry points. One such entry point is the geography of Terezín. EHRI partners created a comprehensive gazetteer of locations in the concentration camp that describes barracks and crematorium, etc. The research guide links these locations to the archival evidence supporting them. There are also other entry points, the guides provide; for instance, via a registry of Terezín prisoners. Archival documentation can enhance the guides dynamically and is itself contextualized with narrative elements such as short biographies of key people involved with the concentration camp or chronologies of events and descriptions (Figure 3).

6. Conclusion

Integrating dispersed and fragmented archives of relevance to a specialised area of research, in this case of research into the Holocaust, has the potential to stimulate new topics of enquiry and generate new research questions. The establishment of trans-national research infrastructures and their integrating activities, as described in this paper, will provide a platform on which researchers can integrate their research across archives and discover new material currently hidden. It can be based on existing research practices, as archival research has always been characterised by the interaction of historians with archival infrastructure. However, at the same time our research demonstrates how many obstacles are still there. This has become clear through our in-depth empirical studies that we could structure around the framework of scholarly primitives and therefore dig deeper into research practices and link them more closely with existing archival service provision.

Our study using scholarly primitives as an underlying framework as well as the subsequent implementation work illustrates that archival research can be enhanced with digital means. Research activities, that are based on gathering, annotating, note-making, etc. and in the process create own rich representation of content in these archives, have to benefit from being able to make these connections virtually. Scholarly primitives can make these connections to research services. We have chosen to implement the EHRI integrated infrastructure with graph databases. With their emphasis on relationships, graph databases are particularly well suited for historical research in particular and humanities research in general.
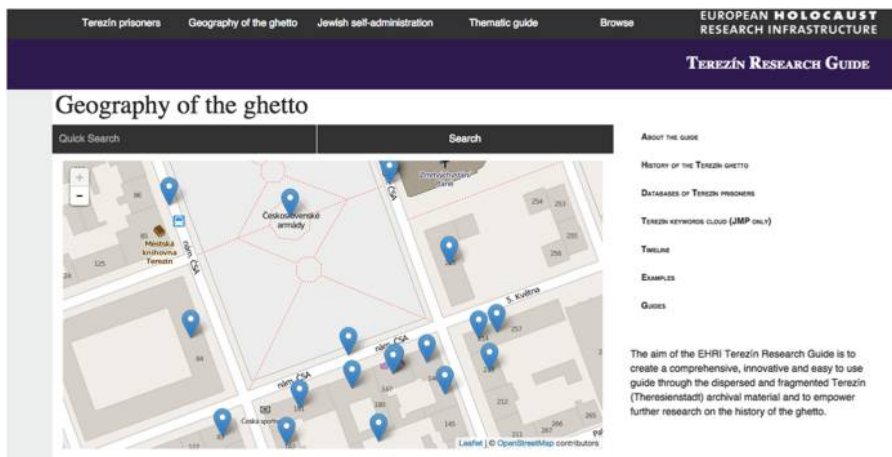

Figure 3.
EHRI trans-national research guide

Based on the graph database approach we could respond to each requirement put forward in our studies. For archival research, we could deliver the required dynamic, research-driven environment, where new material can be integrated and collaboration on content can develop.

Notes

1. http://digitalcuration.blogspot.co.uk/2010/03/linked-data-and-reality.html
2. http://graphbase.net/BlogIntegrationUsingGraph.html
3. www.neo4j.org/
4. http://michaelnielsen.org/blog/the-future-of-science-2

References

Anderson, I.G. (2004), "Are you being served? Historians and the search for primary sources", *Archivaria*, Vol. 58, pp. 81-129.

Anderson, S. and Blanke, T. (2012), "Taking the long view: from e-science humanities to humanities digital ecosystems", *Historical Social Research/Historische Sozialforschung*, Vol. 37 No. 3, pp. 147-164.

Anderson, S.R. and Allen, R.B. (2009), "Envisioning the archival commons", *American Archivist*, Vol. 72 No. 1, pp. 383-400.

Bechhofer, S., Buchan, I., De Roure, D., Missier, P., Ainsworth, J., Bhagat, J., Couch, P., Cruickshank, D., Delderfield, M. and Dunlop, I. (2013), "Why linked data is not enough for scientists", *Future Generation Computer Systems*, Vol. 29 No. 2, pp. 599-611.

Blanke, T. and Hedges, M. (2013), "Scholarly primitives: building institutional infrastructure for humanities e-Science", *Future Generation Computer Systems*, Vol. 29, pp. 654-661.

Blanke, T. and Kristel, C. (2013), "Integrating holocaust researc", *International Journal of Humanities and Arts Computing*, Vol. 7, pp. 41-57.

Blanke, T., Bryant, M. and Hedges, M. (2013), "Back to our data – experiments with NoSQL technologies in the humanities", *Big Data, 2013 IEEE International Conference on, IEEE*, pp. 17-20.

Blanke, T., Bryant, M., Hedges, M., Aschenbrenner, A. and Priddy, M. (2011), "Preparing DARIAH", *E-Science (e-Science), 2011 IEEE 7th International Conference on, IEEE*, pp. 158-165.

Blanke, T., Candela, L., Hedges, M., Priddy, M. and Simenoni, F. (2010), "Deploying general-purpose virtual research environments for humanities research", *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol. 368, pp. 3813-3828.

Bolotenko, G. (1983), "Archivists and historians: keepers of the well", *Archivaria*, Vol. 16, pp. 5-25.

Case, D.O. (1991), "The collection and use of information by some American historians: a study of motives and methods", *The Library Quarterly*, Vol. 61 No. 1, pp. 61-82.

Cook, T. (1984), "From information to knowledge: an intellectual paradigm for archives", *Archivaria*, Vol. 19, pp. 28-49.

Daniels, M.G. and Yakel, E. (2010), "Seek and you may find: successful search in online finding aid systems", *American Archivist*, Vol. 73 No. 2, pp. 535-568.

Dey, I. (2007), "Grounding categories", in Bryant, A. and Charmaz, K. (Eds), *The Sage Handbook of Grounded Theory*, SAGE, Los Angeles, London, pp. 167-190.

Duff, W.M. and Johnson, C.A. (2002), "Accidentally found on purpose: information-seeking behavior of historians in archives", *The Library Quarterly*, Vol. 72 No. 4, pp. 472-496.

Duff, W.M., Monks-Leeson, E. and Galey, A. (2012), "Contexts built and found: a pilot study on the process of archival meaning-making", *Archival Science*, Vol. 12, pp. 69-92.

Dutton, W.H. and Jeffreys, P.W. (2010), *World Wide Research: Reshaping the Sciences and Humanities*, MIT Press, Cambridge, MA.

Edwards, P.N. (2003), "Infrastructure and modernity: force, time, and social organization in the history of sociotechnical systems", in Misa, T.J., Brey, P. and Feenberg, A. (Eds), *Modernity and Technology*, MIT Press, Cambridge, MA.

Edwards, P.N., Jackson, S.J., Chalmers, M.K., Bowker, G.C., Borgman, C.L., Ribes, D., Burton, M. and Calvert, S. (2012), "Knowledge infrastructures: intellectual frameworks and research challenges", report of a workshop sponsored by the National Science Foundation and the Sloan Foundation, University of Michigan School of Information.

Gitelman, L. (2013), *Raw Data is an Oxymoron*, MIT Press, Cambridge, MA.

Glaser, B.G. (2002), "Constructivist grounded theory? Forum qualitative sozialforschung/forum: qualitative", *Social Research*, Vol. 3 No. 3.

Hayles, N.K. (2007), "Intermediation: the pursuit of a vision", *New Literary History*, Vol. 38 No. 1, pp. 99-125.

Johnson, C.A. and Duff, W.M. (2005), "Chatting up the archivist: social capital and the archival researcher", *American Archivist*, Vol. 68 No. 1, pp. 113-129.

Ketelaar, E. (1997), "Can we trust information?", *The International Information & Library Review*, Vol. 29 Nos 3/4, pp. 333-338.

Maharajan, S. (2012), "Performance of native SPARQL query processors", master's thesis, Uppsala University.

Nielsen, M. (2012), *Reinventing Discovery: The New Era of Networked Science*, Princeton University Press, Princeton, NJ.

Palmer, C.L., Teffeau, L.C. and Pirmann, C.M. (2009), "Scholarly information practices in the online environment: themes from the literature and implications for library service development".

Pollock, N. and Williams, R. (2010), "E-infrastructures: how do we know and understand them? Strategic ethnography and the biography of artefacts", *Computer Supported Cooperative Work (CSCW)*, Vol. 19 No. 6, pp. 521-556.

Rockwell, G. (2010), "As transparent as infrastructure: on the research of cyberinfrastructure in the humanities", available at: http://cnx.org/content/m34315/1.2

Rodriquez, K.J., Bryant, M., Blanke, T. and Luszczynska, M. (2012), "Comparison of named entity recognition tools for raw OCR text", *Proceedings of KONVENS*, pp. 410-414.

Ross, S. (2012), "Digital preservation, archival science and methodological foundations for digital libraries", *New Review of Information Networking*, Vol. 17 No. 1, pp. 43-68.

Sadalage, P.J. and Fowler, M. (2012), *NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence*, Pearson Education, Essex.

Schreibman, S., Gradmann, S., Hennicke, S., Blanke, T., Chambers, S., Dunning, A., Gray, J., Lauer, G., Pichler, A. and Renn, J. (2013), "Beyond infrastructure – modelling scholarly research and collaboration", Digital Humanities 2012, Arizona.

Speck, R. and Links, P. (2013), "The missing voice: archivists and infrastructures for humanities research", *International Journal of Humanities and Arts Computing*, Vol. 7 Nos 1/2, pp. 128-146.

Speck, R., Blanke, T., Kristel, C., Frankl, M., Rodriguez, K. and Daelen, V.V. (2014), "The past and the future of holocaust research: from disparate sources to an integrated european holocaust research infrastructure", arXiv preprint arXiv:1405.2407.

Taylor, H.A. (1984), "Information ecology and the archives of the 1980s", *Archivaria*, Vol. 18, pp. 25-37.

Unsworth, J. (2000), "Scholarly primitives: what methods do humanities researchers have in common, and how might our tools reflect this. Humanities computing, formal methods, experimental practice", Symposium, Kings College, London, p. 5.

Weinberger, D. (2011), *Too Big to Know: Rethinking Knowledge Now that the Facts Aren't the Facts, Experts are Everywhere, and the Smartest Person in the Room is the Room*, Basic Books, New York, NY.

Yakel, E. (2004), "Encoded archival description: are finding aids boundary spanners or barriers for users?", *Journal of Archival Organization*, Vol. 2 No. 1, pp. 63-77.

Corresponding author

Dr Tobias Blanke can be contacted at: tobias.blanke@kcl.ac.uk