Intermediating the Human and Digital: Researchers and the European Holocaust Research Infrastructure

Sheila Anderson and Tobias Blanke
King's College London

The research reported on in this working paper was undertaken EHRI Work Packages 16 and 20. The user requirements work (WP16) was led by Sheila Anderson and Reto Speck at King's College London and included The Digital Curation Unit (DCU) at the Athena Research Centre in Athens, Greece, and the Institute for War, Holocaust, and Genocide Studies (NIOD) in Amsterdam. The technical infrastructure work was undertaken at King's College London and led by Tobias Blanke and Michael Bryant. Grateful thanks are due to all our partners for their input into this work.

Introduction

Over the last decade the European Commission has funded the development and building of Research Infrastructures designed to meet the needs of humanities researchers. The first of these, the Digital Research Infrastructure for the Arts and Humanities (DARIAH) is a networked infrastructure of people, tools, digital information, and methods designed to act as a dynamic research ecosystem that facilitates, supports, and helps to grow digitally enabled research. Since the establishment of DARIAH, the Commission has funded a number of discipline specific research infrastructures in the humanities the focus of which is to integrate access to archives, to connect knowledge, and to support the process of research for a particular community of researchers. Among the first to be funded was the European Holocaust Research Infrastructure (EHRI). Launched in October 2010, EHRI aims to provide access to archives, to connect knowledge, and to facilitate and enhance the process of research into the Holocaust. EHRI is developing an open, collaborative research environment which will provide integrated online access to dispersed (archival) resources relating to the Holocaust, initially across Europe and Israel, by linking archives, functionalities, and people.

Researchers researching aspects of the Holocaust face a particular set of challenges as many of the sources they wish to work with are fragmented and partial, partly as a result of the efforts of the Nazis to destroy the evidence of their actions, and partly as a result of the political and nation state realignments that took place at the end of the Second World War. As a consequence researchers may need to work in multiple archives and to gain an understanding of where the records of interest might be; this is complicated by the fact that many archival records are highly dispersed, and at times duplicated, across multiple institutions. The Archives and institutions that hold material of relevance for the study of the Holocaust are wide ranging - many are 'traditional' archives which contain material of relevance to the Holocaust as part of a much wider collection of archives; others were established specifically to record the tragedy and horror of the Holocaust and to remember its victims. Research into the Holocaust is wide ranging and multi-disciplinary with researchers from across a range of disciplines including history, literary studies, sociology, human geography, drama, musicology, and so on covering an extensive range of topics. For example, in Holocaust Studies research into victims'

materials or testimonials stands next to more traditional archival research into perpetrators; researchers may focus on a particular country, Poland or Hungary for example, or on the process of ghettoization, whilst others are researching into aspects of memorialisation, remembrance, and post-war representation of the Holocaust.

EHRI faces a significant challenge then in the 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 sheer range and variety of the research topics and questions addressed by Holocaust scholarship. This complexity requires not only an understanding of how to integrate disparate and dispersed archives but also how to provide a specialized research environment capable of rising to the challenge of meeting the needs of a multitude of 'niche' research areas. In this scenario, one size does not fit all. So where to start in the process of putting together something as complex as a Holocaust Research Infrastructure?

Edmond suggests that research infrastructures should be based on an understanding that "[..] new methods of technologically enhanced investigation in the humanities supplement, rather than supplant, the tried and true methods of close reading, contextualising primary sources within relevant secondary discourse, and contributing to communities of practice consisting of one's peers. It recognises the value of the old methods while seeking also to enhance them" [1]. Anderson has elsewhere argued that research infrastructures should be considered incremental, as arising from the already established procedures and practices of both collection holding institutions and scholars, whilst at the same time enhancing these practices through an interaction with the capabilities of the computer and computational processes [2]. This latter point is crucial. It isn't just a question of understanding researcher practices and processes and then translating these into a digital equivalent; or of taking archival finding aids, catalogues, and indexes and enabling cross-searching across them. It is incumbent upon those of us working in the new digital infrastructures to seek to understand what might be different, what might be new, and to imagine what the effects might be of the shift of analogue practices into a digital world.

Publishing digital finding aids into a research infrastructure that contains finding aids from hundreds of archives is not the same as publishing on a single archive website; integrating archival finding aids might mean breaking apart the carefully constructed hierarchical nature of those aids, or of disrupting the archive's sense of place by creating new archives from across multiple physical archives. Undertaking research work in a research infrastructure is not the same as visiting an archive or searching a single set of archival finding aids; it offers both affordances and constraints, it affords the ability to search across multiple archives and to break free from the constraints of single site searching but it also distances the researcher from the archivist; it offers possibilities for adding annotations and notes but brings to the fore the potential and the pitfalls of sharing those annotations and notes with a wider community of researchers, and with the archives themselves.

Hayles 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 [3] – in this case from the analogue, physical spaces of archives to the digital, distributed spaces of research infrastructures. Hayles argues

that the common understanding is 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' [3]. As Hayles argues, this is true to a point, we have certainly witnessed the evolution of certain practices from the analogue realm to the digital realm – letters to email for example, and books to e-books – where the goal is the same – communication or reading in our examples – but the difference lies in the medium of reception – paper to computer and book to kindle. However, if we look deeper we may also see that there are some more fundamental changes happening – we write far more emails than we ever wrote letters and our language has changed from highly formalised to a greater level of informality. The introduction of the computer effects a shift in behaviour and has thus become a 'fundamental component' [3] of that behaviour.

The formation of EHRI as a collaboration between researchers, archivists, and e-scientists offers a powerful platform through which to explore the intermediation between archival practices, scholarly practices, and the insertion of computational possibilities. Whilst the goal for humanities research infrastructures remains to integrate disparate and dispersed primary sources, information and knowledge collected, created and managed by archives 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. This is both a technical and a social challenge. The key to success for EHRI will be to ensure the development of a socio-technical solution that integrates content and knowledge, is responsive to the variable nature of the content to be integrated, and that remains dynamic in response to changing research practices and new areas of research enquiry as these are afforded by the application of computational technologies. To enable this we must start with a thorough understanding of how and why research is undertaken and the epistemic practices and interactions that support research work using archival sources.

This paper reports on the work undertaken by EHRI work package 16 to investigate the practices and processes of Holocaust researchers and their requirements for the Holocaust Infrastructure and our work to translate these into a flexible technical infrastructure. We first present the methodology used to collect data on practices, followed by a discussion of our approach to the analysis of the data gathered. Comprehensive details of the results of our analysis are provided in EHRI deliverable 16.4 and in two recently published papers [4] and [5]. In this paper we summarise the critical overarching trends that emerged from our research and that are essential in informing our thinking about the technical infrastructure. We finish with an introduction to the system foundations that are being put in place for EHRI as developed by WP20.

Methodology

The first stage of our research to identify and understand research practices in the field of Holocaust studies began with identifying and seeking to understand the Holocaust research landscape through a literature review and annotated bibliography, and an analysis of the bibliography in order to draw out the range and variety of the Holocaust research communities.

To gain a broad understanding of the practices of researchers working on the Holocaust and the process of research and knowledge making in archives we took a four-pronged approach: desk research

synthesizing a wide range of studies on researcher behaviours and practices, knowledge organisation, and information systems development, and scholarly works which described the methodology employed in the process of humanities research in general; 8 semi-structured interviews with archivists and librarians working in a variety of collection holding institutions; 15 semi-structured interviews with researchers; and an on-line survey for Holocaust researchers undertaken over an eight-month period with a total of 277 responses from researchers. 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 as they arose during the interview.

Interviews with archivists were obtained from EHRI partners who hold significant collections of relevance for research on the Holocaust. Requests were sent to 12 archivists and 8 responded to the request. Interviews lasted between 50 and 80 minutes and were digitally recorded, and then transcribed and encoded using the Nvivo software suite. The interviews were structured around the following themes: Information about readers including research background; details of collections and finding aids, including primary collections relating to the Holocaust; the range of support services provided pre, during, and post visit; usage of these services including known gaps and problems; their perception of reader behaviour and how aware and knowledgeable readers are of finding aids; and levels of meditation archival staff perceive themselves to provide to collections and the influence of this meditation; their attitude towards user-created content, crowd sourcing, social media tools. The interviewers also probed the expectations archivists had of EHRI and digital innovation in general and the challenges of working in a large infrastructure.

Potential interviewees for the researcher interviews were obtained from the bibliographic analysis and from personal recommendations from the EHRI research partners. 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, ranging from history, to sociology, languages and literature, archival sciences and visual arts, and to represent different career stages, ranging from doctoral candidates to full professors. Interviews lasted between 45 minutes to 2 hours and were digitally recorded, transcribed, and encoded and analysed using Nvivo. The interviews were structured around the following themes: research area and topics of interest; resources found, appraised and used; use of existing domain literature; organisation and annotation of found information; language; communication, how and with whom; collaboration; workplace; the use and value of digital resources and finding aids; access and privacy; any other comments. The survey questionnaire was designed to complement the interviews and to provide an evidence-based statistical descriptive record on scholarly practice and needs. 277 valid responses where collected between October 2011 and May 2012. Most respondents lived in Europe, while 14% lived in the United States and 6% in Israel. The questionnaire comprised seventeen 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.

Taken together the interview and survey data provide a rich seam of evidence for analysis from which we have abstracted a range of basic activities, a model of the research process and underpinning

interactions, and the defining features of archival research in Holocaust studies. In the next section we provide the conceptual and theoretical basis for our analysis.

Primitives and Processes

If viewed at an appropriate level of abstraction, working practices, even in as broad a domain as the Humanities, tend to involve a finite set of fundamental processes common across disciplines. The concept of scholarly primitives 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 systems design and functionality. In 2002 John Unsworth 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" [6]. Since then, perhaps the most widely quoted synthesis and discussion of primitives is that provided by Palmer and colleagues [7]. Palmer et al 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. Within these overarching categories a further twenty primitives are identified which provide a more detailed level of understanding. For researchers in the humanities the most common primitives were chaining, browsing, collecting, re-reading, assembling, consulting, and note taking [7].

Hedges and Blanke used the scholarly primitives provided by Palmer as a conceptual framework for developing an infrastructure for humanities research work, primarily in the discipline of classics, that is 'sufficiently generic to cater for different research needs while not being engineered beyond the requirements and understanding of researchers' [8]. 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. In our work for EHRI we found five categories to be of relevance for our purposes, and within those categories fourteen primitives to be of relevance as listed in figure 1 below:

Searching: Direct searching, Chaining, Browsing, Probing, Accessing	**Collecting** Gathering Organising
Reading Scanning Assessing Re-reading	**Collaborating** Networking Consulting
Cross-cutting Primitives Note-taking Translating	

Figure 1: EHRI Primitives adapted from Palmer et al.

The process of searching is, of course, fundamental to archival research work, and also fundamental to finding secondary published sources which inform and shape the construction of new knowledge. However, Duff et al suggest that instead of assigning this behaviour as 'searching' our understanding would be enhanced were we to regard it as an essential part of researcher meaning-making, and as a component of the interpretive process rather than 'seeking' activity [9]. The consequence of this approach is that searching cannot be viewed as an activity separate from the processes of reading, note-taking, and creating texts. Knowledge work is thus better conceptualised as a process of creating associations between searched for records and sources as an integral part of meaning-making. Rather than see scholarly activities and primitives as a set of discrete activities to be supported by the EHRI infrastructure we should instead view them as a *process* to be supported.

Unlike Hedges and Blanke we included reading as a core activity alongside scanning, assessing, and re-reading as the underlying primitives. We reinstated reading as the process of reading, assessing, and re-reading proved an essential component of research. Making judgements on the likelihood of a source being of value to answer a particular research question involves a process of reading and judgement before deeper reading can occur. This fits with Duff's argument that the activities undertaken as part of archival research work are an essential part of the meaning-making process. In this instance 'reading' is essentially 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.

We also included the primitives added by Blanke and Hedge of *comparing* and *delivering* as essential components of the infrastructure. The investigative process of research outlined in the next section indicates that comparing records and sources, choosing some, and discarding others, comparing the relevance and the information contained within is a key element of the research process. The activity of 'delivering' as conceptualised by Hedges and Blanke assists us to think about what is to be delivered, how, and under what constraints. For example, what rights should be associated with sharing, or to limit the extent to which sharing takes places, either by limiting the content to be shared, or by limiting who the content can be shared with.

We thus arrived at the following set of activities and primitives to guide the EHRI infrastructure:

Searching *(investigating)*	**Collecting**
Direct searching,	Gathering
Chaining,	Organising
Browsing,	
Probing,	
Accessing	
Reading (*interpretation)*	**Collaborating**
Scanning	Networking
Assessing	Consulting
Re-reading	*Sharing*
Cross-cutting Primitives	***Comparing***

Note-taking Translating	***Delivering***

Figure 2: Palmer's Scholarly Activities and Primitives extended for EHRI purposes: EHRI Activity additions in bold italics and EHRI primitives in italics

However, modelling practices solely in the context of primitives provides only a partial understanding of the 'dense forest' [10] of knowledge making practices. Whilst the concept of primitives facilitates the identification, description, and prioritization of the activities undertaken during research work, it fails to explain the process and the interdependencies that are inherent in research work. There is a need for an explicit common conceptual and representational framework within which these proposed sets of activities could be systematically compared, extended and, perhaps, jointly exploited. The Digital Curation Unit's Scholarly Research Activity Model (SRAM) inspired by activity theory, the standard (ISO 21127) ontology for cultural documentation CIDOC CRM and conceptual models of business processes, provides such a framework and also enables systematically representing the typology and structure of activities, associating activities to actors and resources, and maintaining "purpose trails", i.e. associating activities with particular questions and research goals at various levels of generality [11].

Within SRAM the concept of *research goal* is instrumental, since it enables capturing the successive refinement from high level objectives down to narrower goals and concrete questions, and places these in the context of the overall research process to be supported by the infrastructure [REF]. Using SRAM allows questions to be asked such as: what was done, where, why, and when, and in what steps? Who did what and why, using which method? How was each step performed and what were the overarching goals driving the work forward? It also facilitates the identification and analysis of the process of research which, following our analysis of the Scholarly Activity and Primitives model, emerged as essential. At the heart of SRAM is the researcher with a research goal, that is, a research topic to investigate and research questions to seek answers for. The concept of research goal allows the research to be set in the context of other research in the area, including connecting it to published works in the field and locating it within the underlying theories, ideas and concepts that are informing the research.

The SRAM provides a framework to take scholarly activities and primitives and to render them dynamic by asking how researchers search and find resources, how the activity might change over the course of research, what the relationships and interactions are that support these activities, and how we can understand their nature. We are able to gather information on the history and practices of the content providers, in the case of EHRI for example, to understand the policies of a collection holding institution and in what way these might influence the way in which it acquires and manages its holdings. We can specify the nature and form of the information sources in their care - their format for example, or whether there are lacuna, and what tools and services are available to support knowledge work. We are also able to gather information on the tools and methods employed by researchers and to compare these to those provided by archives and other collection holding institutions. SRAM provides us with the necessary tools to start to unpick the complex web of people, activities, and interactions, and to imagine how this might translate to the infrastructure.

Taken together the concept of primitives and the Scholarly Research Activity Model provide a powerful set of tools with which to understand the intermediation between human research practices, archives, and digital infrastructures, and to imagine the effects of the shift from one medium to another.

Research as an Investigative and Social Process

The analysis of our data using both primitives and SRAM leads us to assert that scholarly research is deeply embedded in a complex network of relationships and interactions that includes other researchers, archivists, knowledge of secondary literature, knowledge of archives, and knowledge of sources. Researchers make continuous use of earlier results both to inform current research and as instruments to identify areas for further investigation: research starts with what other researchers have done, the sources they have used, the methods employed, and the tools used.

Latour [12] has demonstrated that social context is essential to understanding scientific activity. Research work takes place within, and is informed by, the normative practices and expectations held by a particular domain and the activities and processes of research arise from and are informed by the cognitive models shared amongst domain communities. Our desk research into the methods and approaches of humanities scholars indicates that the humanities exhibit a particular set of features:

- Hermeneutic in nature rather than experimental, and not seeking formal laws and explanations
- Primarily based on narrative, text, and rhetorical in nature
- Recursive, not linear – a constantly questioning process
- Primarily deep reading / reasoning of sources - there are, of course, some exceptions where researchers are using large quantities of data and quantitative methods, but it would be fair to say this is the minority, even if a growing one
- Individualistic, even when collaborating – a sense of uniqueness is essential
- Sources rather than 'data' – the human record including texts of all kinds, images of all kinds, objects, artefacts, video, film, audio (speech, music)
- Based on existing intelligence and knowledge expressed in publications and grey literature – these are the starting point and a source of both inspiration and knowledge
- Complex, varied, multi-faceted and increasingly framed around a mix of digital and analogue sources and processes, and moving between digital media and physical spaces as the place for research

The hermeneutic nature of research is reflected in how archives are approached. Our research confirms that scholarly work in archives is complex and multi-dimensional with sources selected from across a range of archives, combined, and used to extract knowledge. It is not a linear process but is recursive and interpretive. The ability to find and assess these sources is highly variable, some are easy to locate whilst others require a long and detailed investigative process. Researchers move beyond

'searching' as defined by primitives and instead investigate and track down sources that may be of interest, following hunches, and being open to serendipitous discovery. The 'search' process is, as Duff suggests, part of the meaning making process with interpretation and assessment of the sources included from the start of the process. At each stage of this process of interpretation traces are discovered and discarded, created and noted, until such time as the work coalesces into a wider understanding of the topic under consideration. The following quote is from an extended case study produced by Anderson in collaboration with a post-doctoral researcher:

> "I think a better term than 'search process' 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. Over the course of my project, I evolved a strategy for dealing with archives which included initial scouting and getting an overview of archival contents and possible leads (either online or in person), then in-depth reading and/or investigating of target areas identified, then following-up of side leads and tangents, then less targeted browsing of online catalogues and finding aids, which might in turn throw up new leads". (Case Study 1, Literary Scholar)

This investigative process extends to working across multiple archives and using multiple sources in the process of their research:

> "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. I'm also using other types of sources like press material, diaries, testimonies, personal collections and also oral history and of course I'm using images including not only photographs but also posters, leaflets, other visual images and film footages." (Interview with Holocaust Researcher)

> "My sources are: secondary literature, we should include all published testimonies around the Holocaust in Greece, meaning memories of survivors, the few books that have been written and effectively transform the individual experience in historical writing. Besides secondary literature, sporadic and scattered archival sources, archival material from the General State Archives, that contains mainly court papers from the early 60's and specifically the way of war reparations claim and payment. The Historic and Diplomatic Archives of the Ministry of Foreign Affairs is very rich…… As far as full texts of Jewish archives I'd say the archive of the Jewish Properties Management Service, and organization that was created in '43 to arrange the collaboration with the Germans about what will happen with the movable and immovable property of the deported Jews, and the archive of the Association for the Healthcare and Rehabilitation of Israelis of Greece. Both these archives are held by the Central Jewish Council." (Interview with Holocaust Researcher)

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. Given the variable nature of the descriptive records created by archives, this is a difficult problem to address and one that must also take account of the search strategies of Holocaust researchers. 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 search process is complicated by the nature of archival finding aids. Archival finding aids are constructed to assist the management of archives and are designed to reflect the original order and provenance of the records. As such, they rarely provide the topic based approach preferred by researchers. Researchers' questions tend towards the conceptual, framed around a topic rather than a hypothesis, for example, the selling of Jewish goods, Jewish resistance, or the role of music in the Concentration Camps. There was some divergence of opinion concerning the usefulness or otherwise of archival findings aids. 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. A minority expressed at least some difficulty in using finding aids or in the lack of finding aids:

> '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. You have to go in every day and build a relationship so that they know that you are someone who's serious and then to try and work with what's there. I think I'm using a very difficult place to work." (Interview with Holocaust Researcher)

What was common to many however, was the establishment of a good relationship with the archivist with many expressing 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. In this respect developing a good relationship with the archivist was seen as vital:

> "The personality of the archivist is very important. If the archivist is not helpful that might be a nightmare" (Interview with Holocaust Researcher)

Archival Research Guides were also considered as useful tools for the researcher:

> "Particularly useful was Colin Thom's Researching London's Houses: An Archives Guide (2005), which gave me precise instructions on what resources the London Metropolitan Archives held for research into buildings, occupants, and street plans". (Case Study 2, Historical Geographer)

The archivist and the archival research guides as sources of advice on the wide landscape of records and evidence that is available in other institutions, and in many cases for Holocaust research, in other countries, is valuable particularly to more inexperienced researchers just embarking on serious archival research.

However, the archivist researcher relationship is not always straightforward and the role of the archivist unambiguous as a source of help and support. Archivists emphasised their role as experts who support researchers and provide context for the sources but some researchers disagreed with this assessment suggesting instead that archivists could actively discourage with their reluctance to help:

> "They have suspicions about anyone but especially foreigners so it is hard to work there. I think that in […] archivists don't always want you to look at the material" (Interview with Holocaust Researcher).

Concern was also expressed about access to archives and in particular about access to private archives. Researchers of the Holocaust in Eastern Europe and Greece identified this as a particular problem and suggested a need for EHRI to provide thematic descriptions of privately held material regardless of the access possibilities to the material itself.

Interviews with archivists on the other hand indicate that this conceptual framing of research questions does not translate well into the classification schemes and finding aids provided by archives. The consequence, according to the archivists, is that researchers frequently rely on archivists to help them with their initial enquiries and to instruct them on the best way to use the finding aids on offer. This perception may be a consequence of who approaches the archivist. The more experienced researcher is less likely to request help while the less experienced is more likely to request help [13].

This is discussed in more detail in [4], where the conclusion is that archivists and reference librarians can play a crucial role in the process of finding-out-about, using their wide knowledge of the field to direct researchers to sources they may not have considered, and in many cases by directing researchers to archives other than their own which may hold sources of interest. The 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 [4]. This resonates with the findings of Yakel [13], who suggests that users find navigating through digital Encoded Archival Descriptions (EAD) finding aids problematic with users failing to understand how records are represented in the EAD hierarchy, baffled by the jargon, and frequently getting 'lost' as they searched for information. Despite these draw- backs, digital finding aids remain an important access route to locate and access sources of interest as part of an overall strategy of personal and digital forms of communication and investigation.

This is a point emphasised by others. Terry Cook suggests that archivists should 'document function, activity, and ideas, rather than primarily reflecting the structures, offices, and persons of origin'. Cook also argues that archivists should be open to researchers' insights and include researcher annotations in their finding aids [14] A similar case is made by Blouin and Rosenberg [15] who suggest the construction of 'parallel but linked' access with scholars creating parallel access systems linked to archival finding

aids that reflect a historical understanding of the sources. At its most basic level this could include keywords defined by researchers assigned to finding aids through to a sophisticated system that included bibliographies, citations, details of research projects, and so on. Including the traces created during the process of research has the potential to enrich finding aids and to provide alternative pathways to navigate through the myriad of sources and archives. EHRI has an opportunity to not only integrate archives and sources and provide cross-searching of archival descriptions, but also to enrich these with the addition of researcher-generated terms and classifications, and to encourage and facilitate 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 process might also be taken a step further. The introduction of computers and computational processes into an infrastructure offers the potential to leverage access to the outputs of the process of research. During research 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 and in the process build up 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 very complicated process:

> '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. I also have many files that I have transcribed for my dissertation: wills, marriage contracts etc. These are archived thematically and inside chronologically, the title of the envelope is marriage contracts and inside chronologically. 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 two fields with keywords, (family) or dates (middlewar, 1940), a field where I keep a note if the book is good, if I have read it and another field where I keep comments, about many things, like where it is, if I have it, if it is here. I also keep there articles and magazines. 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; there is one database per person, like a personal folder for each person, that was made by processing some folders and the data I've found here, name, birth name, date of birth, education, date of arrival, date of departure, in which assignments was he a part of, 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)

Archivists recognize this alternative knowledge building, and indeed, they use it to enhance their own knowledge of the archives in their care.

> "Annotations is (sic) 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 rarely is this knowledge formalized to enhance finding aids; instead it remains in the head and on the computer and notebooks of the researcher and is only partially published in articles and books as the outputs of research. To utilise this material requires both willingness on behalf of researchers to share the outputs of the process of their research, something at present primarily only undertaken informally through conversations and at conferences, and for archivists to be willing to open their finding aids and to be receptive to multiple representations of the archives in their care.

> "[..] 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)

> "When I am starting to work on a new topic I speak with other historians about that topic if they have any ideas about it because the other people don't think like you think, so they will think about things you never think about. I find that very useful. They will probably tell you something very useful, to look this way although you may never have thought about it. So 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 I ask people to read me" (interview with Holocaust Researcher)

Our research indicates that overall both archivists and researchers believe that enhancing, adding, and annotating metadata is important but in many cases with the caveat that control mechanisms are in place. For example, one of the archivists interviewed 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 state they would be prepared to share resources and information on their own work with others. One of our interviewees illustrated how he already worked with a librarian 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 deli.cious and so on" (interview with Holocaust researcher)

We conclude from this that EHRI should provide not only a Portal though which researchers can find, retrieve and use EHRI content, but should also provide a social space in which EHRI researchers can share and exchange knowledge, expertise and information; in which researchers and archivists can communicate and exchange knowledge and information each enriching the perspective and offerings of the other; and in which archivists might also share knowledge and information about their dispersed collections and seek to link and connect not only their finding aids but also at deeper level of granularity. In short, EHRI should aim to operate as a digital research ecosystem [16].

Research Infrastructure as Intermediary

Following the model of an ecosystem, we have aimed for system foundations that allow an environment to grow dynamically both in a technical and social way, but that is at the same time efficient and quick to access. In what follows, we present the principles of that design, which we present in more detail in [17]. 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 and develop an environment that was 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, as described earlier. We cannot describe all the components of this infrastructure here, but concentrate on those that fulfil key requirements discussed above. One of the main components the EHRI infrastructure will provide is an integrated research environment for Holocaust researchers to work through the data sets and collection descriptions gathered by the project. Researcher questions will change depending on the material, while at the same time they need the material easily accessible and delivered to them in a satisfying amount of time. This has become clear not just in the user requirement work that we have undertaken for EHRI but also in other work we have done in the field of arts and humanities e-Research with different kind of user groups [18] and [19].

The requirement of a dynamically evolving research environment can be seen not just for arts and humanities research but across several other disciplines where the focus is on interacting with the research data. In fact, this focus on the ability to organise, research, analyse, and create exchanges around data has become one of the few general commonalities that has bound together disciplines in the various e-Science programmes. In 2008, Dave de Roure [20], one of the pioneers of the UK e-Science programme, consequently announced a new e-Science in a keynote at the annual IEEE e-Science meeting, which concentrated on enabling this new kind of interaction which data intensive research requires. We would like to suggest that the new e-Science needs to have a second look at its fundamental data integration work and approaches to enable this flexible interaction and collaboration with data. We need data work and infrastructures that that can change with the new questions that e-Science wants to enable without losing data or making complicated changes to its formats and schemas.

For EHRI, the integration of Holocaust material means understanding first of all the landscape of the archival institutions, organisations and practices involved. In particular, we needed to not just understand the research practices that take place within these archives but also needed to take into account the specifics of the collections. In the confusion of the Second World War and its aftermath, exacerbated by the perpetrators' attempts to destroy the evidence, collections of Holocaust material are especially dispersed and difficult to access. But, it is these currently unknown documents in potentially hidden archive that interest a historian of the Holocaust looking for new answers to existing questions or new questions altogether. As discussed, this is one point where the requirements for research infrastructures for archives differ from other archival integration projects, which are often more concerned with integrating existing publicly available material to develop an integrated catalogue.

This means that we cannot foresee what kind of data we need to integrate next and that realistically we must expect it to be unique in its setup and formats. As discussed above, Holocaust will have to work with a great variety of sources and our infrastructure needs to enable this.

One solution that we did not want to follow is to limit the amount of information we provide on some common fields we assume to be found in all collection holding institutions. This will be unacceptable to researchers as it limits the amount of information they have access to. Our user evaluation work has shown that archivists' knowledge about their domain is key to the success of research infrastructures. As discussed, they are currently the gate keepers that help a researcher dig deeper into the data. The 'archivist's head' provides context to the research needs. Secondly, it has shown that the 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. This is the background of the essential research primitives of note-taking and annotations. We started to investigate a new kind of infrastructure that would allow us to focus on the content as a researcher would need it and take the data as we found it, heterogeneous and often incomplete, from different sources but even in this form useful to the researcher. Furthermore, the environment needed to support ad-hoc queries against the collected material and queries we might not yet have identified that were based on connections we could have not yet anticipated. Both requirements do not match well to traditional data stores such as relational databases, which often require a predefined understanding of the underlying data structure and which often require redefining this data structure should there be a new type of connection.

Alternative semantic web approaches, while more compatible with ours, still to a large extend rely on triple stores queried via SPARQL to investigate this data and are therefore not flexible enough for us either. This model has proven to be difficult to understand not just for researchers but often for developers, too. Since SPARQL lots of research has gone into designing and developing more effective search and browse environments based on semantic web approaches. Our own research has shown [21], that especially the graph structure of RDF suits the typical ways of how humanities researchers would like to explore sources. They support the traditional types of browsing through connected sources, as described above. But, there are further disadvantages to the use of semantic web technologies like triple stores that make it less suited to the data-driven research we would like to enable. We follow the critique of Saunderson [22], who convincingly presented how RDF and triple stores pay too little attention to the data and too much to the structure. He recommended investigating new NoSQL technologies instead.

Our work indeed shows that NoSQL technologies can support a data infrastructure that supports the kind of socio- technical environment that we are seeking, and they seem to be an excellent building block for a research ecosystem. In the remainder of this paper we will concentrate on the part of the EHRI infrastructure that uses graph databases, a specific kind of NoSQL technologies that also allows us to ingrate all the many advantages of semantic web approaches for the publication and consumption of resources, as they can also function as a SPARQL endpoint [23]. Graph databases have offered us a new approach that sup- ports deep and rich investigation of data and seem a natural fit to

our application domain of a research-led archival integration. They are a relatively old technology [24] that has come to new prominence and achieved a new level of maturity within the NoSQL family of data stores. While most of the NoSQL databases are primarily concerned with the challenges of big data, graph databases address a different challenge traditional relational databases do not address well; a larger number of smaller records that are, however, heavily interconnected [24]. We have chosen graph databases rather than other types of NoSQL, as they can grow easily with any kind of new information that is added to them. According to [24], they scale best towards complexity or towards information that is not uniform. Secondly, they 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 we have argued elsewhere in detail [20], most humanities data possesses a complex structure, with many internal relationships both structural and semantic.

We use the Neo4J graph database [24]. Its data model has three simple elements: nodes/vertices (V), edges that represent relationships between nodes (E) and properties that can be assigned either to nodes or relationships (λ): $G = (V, E, \lambda)$. In the course of our investigations, we found that our source components can be mapped onto this model. Archival descriptions can be nodes and connected via edges with other descriptions or their repositories, researcher annotations are nodes that can be linked to these descriptions and the material they refer to, and typical archival thesaurus terms are again nodes where the edges between them represent the typical narrower and broader relationships between thesaurus terms. As long as it can be mapped onto the node relationship model, any new data model can be added to the existing data. New data sets can be added to this environment by mapping their data model onto nodes and edges between them, which means they scale towards information that is not uniform and the way that researchers combine sources with concepts, using their own mark-up and notes. Furthermore, different data domains can be directly added. This not only applies to thesauri added directly to the sources but also to social data from researchers' LinkedIn or Facebook profiles, especially since Facebook promoted 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 [26], providing links between sample researchers' and archivsts' 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 the real world 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'. As we also implement time stamps for all our graph activities, this can even be linked to particular time slots in order to reflect the above discussed dynamic process of reflection inherent in humanities research.

Graph databases assume that most of the enquiries of the data stored in them will follow the relationships between these data items. With triple stores, any new added relationship will have an impact on the overall performance of SPARQL queries as they rely on graph matching techniques against

the whole data set. The Neo4J graph database we use, on the contrary, is based on exploring the data by traversing the data from a localised point of entry. Further technical details are described in [18], where we also present how we use an external SOLR that stores the textual and linguistic properties of archival description data index to retrieve relevant collections for the portal. Retriev- ing the full context of the documentary unit can then be achieved by walking through the graph from there. In terms of computational complexity this traversal is constant (O(N)) and independent of any new relationship that has been added to the graph outside the context of the documentary unit. Neo4j also comes with a range of more complex graph query languages such as Gremlin and Cypher as well as integration with more advanced graph analysis frameworks such as JUNG [25] that allow for the efficient processing of more complex graph operations such as a shortest path analysis, etc. We are currently testing these with researchers and our results will feed into developments of the infrastructure.

Conclusion

Selecting and 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 large research infrastructures and their integrating activities as described in this paper will provide a digital platform on which researchers can follow an investigative trail across archives, connect and link dispersed evidence, and discover new material currently hidden away in the miasma of a large archive series. However, our research demonstrates that whilst providing access to and integrating archival finding aids is, of course, an essential part of enabling new research, it is not sufficient to fully achieve the transformative potential of integrating activities.

Our study illustrates that in order to use archives effectively researchers form complex networks of support including seeking advice and guidance from other researchers and from archivists. During the process of research, researchers gather, annotate, and make notes on the archives, creating in the process rich personal collections of archival materials. Little, if any, of this information is fed back into the finding aids. Research infrastructures then, are best theorised as intermediating ecosystems that recognise these complex networks and results in a technical architecture that supports this complexity. We have chosen to implement the EHRI integrated information resource using graph databases and to experiment with social media. With their emphasis on relationships, graph databases are particularly well suited for historical research in particular and humanities research in general. The next steps are to continue our work exploring the potential of graph databases, and to investigate further the social platform for EHRI to support communication between researchers, and between archivist and researcher.

References

[1] J. Edmond (2013) “CENDARI’s Grand Challenges: Building, Contextualising and Sustaining a New Knowledge Infrastructure” International Journal of Humanities and Arts Computing 7(1-2): 58-69.

[2] S. Anderson (2013) “What are research infrastructures?” International Journal of Humanities and Arts Computing 7(1-2): 4-23.

[3] N. Katherine Hayles (2007) “Intermediation: The Pursuit of a Vision” New Literary History 38.1: 99-125

[4] R. Speck and P. Links (2013) “The Missing Voice: Archivists and Infrastructures for Humanities Research” International Journal of Humanities and Arts Computing 7(1-2): 128-146

[5] A. Bernadou, P. Constantopoulos and C. Dallas (2013) “An Approach to Analyzing Working Practices of Research Communities in the Humanities” International Journal of Humanities and Arts Computing 7(1-2): 105-127

[6] J. Unsworth, (2000) “Scholarly primitives: what methods do humanities re- searchers have in common, and how might our tools reflect this,” in Humanities Computing, Formal Methods, Experimental Practice Symposium, Kings College, London, 2000

[7] C. L. Palmer, L. C. Teffeau, and C. M. Pirmann (2009) “Scholarly information practices in the online environment,” Report commissioned by OCLC Research. Published online at: www. oclc. org/programs/publications/reports/2009-02. PDF.

[8] T. Blanke and M. Hedges (2011) “Scholarly Primitives: Building Institutional Infrastructure for Humanities e-Science” Future Generation Computer Systems, doi:10.10.16/j.future.2011.06.006

[9] W.M. Duff, E. Monks-Leeson and A. Galey (2012) “Contexts Built and Found: a pilot study on the process of archival meaning-making” Archival Science 12:69-92

[10] C. Camic, N. Gross, and M. Lamont (2012) Social knowledge in the making. University of Chicago Press

[11] A. Benardou, P. Constantopoulos, C. Dallas and D. Gavrilis, “A conceptual model for scholarly research activity,” iConference 2010 Proceedings, pp. 26–32, 2010

[12] B. Latour (2005) Reassembling the social-an introduction to actor-network- theory. Oxford University Press, ISBN-10: 0199256047.

[13] E. Yakel (2004) “Encoded archival description: Are finding aids boundary spanners or barriers for users?” Journal of archival organization, vol. 2, no. 1-2, pp. 63–77

[14] T. Cook (2011) “The archive (s) is a foreign country: Historians, archivists, and the changing archival landscape,” American Archivist, vol. 74, no. 2, pp. 600–632

[15] F. X. Blouin and W. G. Rosenberg (2011) Processing the Past Contesting Authorities in History and the Archives. Oxford University Press.

[16] S. Anderson and T. Blanke (2012) “Taking the long view: from e-science humanities to humanities digital ecosystems”, Historical Social Research 37: 3 pp. 147-164;

[17] T. Blanke and C. Kristel (2013) “Integrating holocaust research,” International Journal of Arts and Humanities Computing, 7(1-2): 41-57

[18] M. Jackson, M. Antonioletti, T. Blanke, G. Bodard, M. Hedges, A. Hume, and S. Rajbhandari (2009) "Building bridges between islands of data — an investigation into distributed data management in the humanities," in Proceedings of the Fifth IEEE International Conference on e-Science. Washington, DC, USA: IEEE Computer Society

[19] T. Blanke, L. Candela, M. Hedges, M. Priddy, and F. Simeoni (2010) "De- ploying General-Purpose Virtual Research Environments for Humanities Research," Philosophical Transactions of the Royal Society A, vol. 368, no. 1925, pp. 3813–3828

[20] D. de Roure (2008) "The new e-science," http://www.slideshare.net/dder/the-new- science-bangalore- edition

[21] T. Blanke, G. Bodard, M. Bryant, S. Dunn, M. Hedges, M. Jackson, and D. Scott (2012) "Linked data for humanities research — the spqr experiment," in Digital Ecosystems Technologies (DEST), 6th IEEE International Conference on Digital Ecosystems

[22] R. Sanderson (2013) "Rdf: Resource description failures and linked data letdowns," http://www.cni.org/wp-content/uploads/2013/04/CNI RDF Sanderson.pdf.

[23] P. J. Sadalage and M. Fowler (2012) NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence. Addison-Wesley Professional

[24] M. Rodriguez (2013) "On graph computing," http://markorodriguez.com/2013/01/09/on-graph-computing/

[25] neo4j (2013) "Fun with facebook in neo4j," http://blog.neo4j.org/2013/06/