Integrating Holocaust Research

Tobias Blanke and Conny Kristel

Introduction

In 2010, the European Commission decided to fund a new initiative in the domain of Holocaust Research, the European Holocaust Research Infrastructure (EHRI). For the first time, European research money was put into Holocaust research with a focus on developing a sustainable research infrastructure rather than funding only individual projects. Prior to this, this area of research funding was the domain of foundations such as the Claims Conference, etc. But, maybe even more remarkable was the programme EHRI was funded under. The FP 7 programme of Integrating Activities aims at providing links between existing research infrastructures on a local and national level and at building bridges between these. Before this, there were other initiatives for humanities research infrastructures but none was funded as an FP7 Integrating Activity. At that time, the only template EHRI had for the proposal were suggestions for large scientific infrastructures that provided trans-national access to telescopes for astrophysics or laser laboratories for material research. The challenge was to map our ideas for an integrated platform for Holocaust research with these existing examples of Integrating Activities. We found that this was much easier than we had initially thought and that the idea of Integrating Activities, as formulated in the FP7 programme announcements, fits Holocaust research that has the ambition to be trans-national very well. [1]

This paper will discuss our ideas and processes for a Holocaust research Integrating Activity and will present the new research opportunities that stem from them. We begin our discussions with an example that demonstrates the need for a research infrastructure and continue with an analysis of our attempts to support the physical integration of Holocaust material with travel grants and fellowship. We finally present our new approach taken to integrate Holocaust material virtually in a portal and digital research environment, which is based on deploying a novel graph-based data integration infrastructure.

Archival research on the Holocaust

Holocaust historians have been working actively with their own traditional infrastructures for a long time without calling them research infrastructures. These traditional infrastructures have been archives. EHRI provides Holocaust historians with access to these, while at the same time opening up new research by offering unprecedented possibilities to link to new sources in a flexible, research-driven manner. Archival work dominates many fields in history.[2] Researchers begin with an orientation on the primary and secondary sources, whereby the overall consensus is that innovative research needs to work with primary sources. In this section we will explain typical challenges in archival research based on an example from Holocaust research.

The challenges in integrating Holocaust material from archives is best explained with an example that shows how much needs to be done. There are many examples, but the early documentation on the ghetto in Terezín (Theresienstadt in German) is here a particularly good case.[3]Hans-Günther Adler (1910-1988) was born in Prague into a German-Jewish family and became a writer at a relatively young age. When he was deported to Terezín in February 1942, he began to document life in the ghetto in order to be able to become a scholarly witness, if he would survive the war. His attempt to document what was going on around him helped him to get through his years in Terezín. Later on, Adler would become one of the early historians of the Holocaust. In 1955 he published a study on Terezín called '*Theresienstadt 1941-1945. Das Anlitz einer Zwangsgemeinschaft*'.[4] Like so many Holocaust related collections, his papers and books have become scattered over a number of countries.

Adler was liberated in a subcamp of Buchenwald in early April 1945. In June he returned to Prague, where he soon started working in the Jewish Museum on assembling a collection on Terezín Ghetto. Leo Baeck, famous German rabbi and scholar, who was also imprisoned in Terezín, had managed to keep the Adler collection together and gave it back to Adler after the liberation. Apart from the Jewish Museum and Adler, there was also an initiative from a group of Zionist activists funded by the Jewish Agency assembling material on Terezín. Most of this material is now part of the rich collections of Yad Vashem in Jerusalem. In February 1947 Adler emigrated to London taking his personal collections with him.[5]

So Adler had left a long trail of documents on Terezín in several countries. Currently, material from the Adler collection is held by at least five different EHRI partners. The Jewish Museum in Prague is an important repository of Terezín-related material. Yad Vashem in Jerusalem administers the documents, which were assembled by the 'Dokumentationsaktion' by the Jewish Agency. Adler's library is in London and now part of the library of King's College London, where Adler's son used to work. The International Tracing Service in Bad Arolsen in Germany was set up to record what happened to civilian victims of the Nazis and holds a registration card of Adler. And maybe most surprisingly, the Dutch NIOD Institute for War, Holocaust and Genocide Studies in Amsterdam has a substantial Adler collection, because of the personal friendship between Adler and the first director of NIOD, Loe de Jong. The latter was also a survivor-historian and had known Adler for a long time. In his will Adler had stated that his collection (largely on Terezín) should be given to NIOD.

Next to the geographic dispersion of the Terezín material, there are also challenges for historical research stemming from the way documentation has been attempted up to now. The Adler collection is also a good example for the state of Holocaust documentation, as part of the documents in the collection are photocopies. While the originals stayed with Adler's son, Jeremy Adler, photocopies of the originals have been done at different times in the past and with different overall aims, which makes it often difficult to understand the context of the documentation. The 'fate' of Adler's collection demonstrates the dispersion of Holocaust related material, which makes it very difficult for researchers to get an overview of the relevant material and the right understanding of its context. A researcher of Terezín will first think of a visit to the Jewish Museum in Prague, and it is also likely that she will contact Yad Vashem. NIOD in Amsterdam or King's College in London will probably be not on her mind at first. There is also no guarantee that not further unknown collections exist elsewhere, which we are not yet aware of and might provide a different view of events.

The dispersion of the Adler/Terezín collections is just one example for the archival reality that Holocaust researchers have to cope with. In recent decades, there have been several smaller scale initiatives, mostly by the national commemorative institute Yad Vashem (Israel) and the United States Holocaust Memorial Museum (USHMM; Washington DC). Supported by the Claims Conference, they have undertaken huge efforts to bring together an impressive amount of documents and collections pertaining to the Holocaust (often in very difficult circumstances, working with local people). Originally, they used to photocopy the originals (and sometimes even existing photocopies), while in a later stage they started scanning documents. As these efforts were focused on material relevant for the study of the Holocaust, they sometimes selected only individual items from collections for photocopying or scanning. Collections with relevant historical material were therefore taken apart.

As a result the scans/photocopies of documents from all kinds of Holocaust material are now physically available in both Washington DC and Jerusalem. In some cases, staff at Yad Vashem and USHMM have prepared their own finding aids, which may differ from the original ones.[6] The representation of the original material has thus also been altered. All of this results in difficulties for historians to judge the evidence in the documentation. Documents are decontextualized and their original provenance might be lost. The construction of an integrated view on the documentation therefore is often a painstaking effort, which requires the overview of the documentation not just in one archive but in many related ones. Historians have to use their own personal networks and become their own archival experts. They must develop their own knowledge of how collections are represented and collected at different places across the world.

Integrating Holocaust archives for research

In a perfect archival world all collections would have been described using the same metadata standards. In this world, researchers have enough information to be able to assess the collection in its context and provenance. But, this perfect world almost never exists, which is not just the result of problems with keeping the collections together after the war or documentation efforts after the war. It is also the result of different practices in the archival and historical sciences which mean that one can often hardly speak of traditional archives as research infrastructures for historical research.

The first research activity in EHRI was to identify existing archives and collections but in such a way that this identification work would directly benefit the researcher. The focus here is especially on those archives that are not part of larger infrastructures and are 'hidden' to most researchers. Here, we found that there is still much more work to do than one might think, not just in parts of Eastern Europe but also elsewhere in Europe, where local archives do not seem to be connected. We have developed an environment to provide high quality descriptions for relevant collections, including the hidden ones. Finally, we provide researchers with travel grants and stipends to visit archives in order to research their collections. From the Terezín case, it is clear that for research into the Holocaust it is very likely that one has to travel, as the research resources are dispersed across different, sometimes surprising locations. Neither copies, nor online finding aids or catalogues are currently a substitute for going through the collection – preferably the physical one – in situ and for building the own historically oriented overview of the collections.

In the remainder of this section we will describe how we improve Holocaust collection descriptions for research, which we will make available online, and on the analogue side, how EHRI provides travel grants for transnational access to existing infrastructures in Holocaust research. Both activities are central to EHRI's ambition to improve the access to Holocaust material and enable new research. Furthermore, support for both kinds of activities is readily available through the up to now science-dominated Integrating Activities programme in Europe, which underlines why this is an excellent research opportunity for humanities researchers. Where scientists need to visit labs and telescopes, where data is led and processed, historians need to visit archives.

In the first phase of EHRI, we have identified over 1,500 Holocaust related institutions and collections that would benefit from research-oriented integrating activities, and national reports have been compiled. They provide a general overview of the history and archival situation in a large number of countries in an easily accessible format. The reports will be published online in 2013. They give a brief general history of World War II (statehood; German rule/influence) and Holocaust history (size of the pre-war Jewish community, estimate of the number of Jewish victims), and secondly a description of the archival situation (organization of archives and legislation) and of institutions that hold Holocaust-related archives. Lastly, the reports provide a concise overview of the state of research of EHRI in the country.

At the time of writing, EHRI focusses as a starting point on a number of so-called key collections. While existing finding aids are our point of departure, ideally EHRI will produce a new collection description, which is researcher-oriented. The traditional way of organizing archival collections, on the contrary, often means that archivists keep the collection together in its original physical context, provide a detailed description of the origin and history of the collection, and a more or less detailed description of the content of the collections. This quality assessment by archivists focuses on the archival context, which is often not the historical context, as no reference is made to the content (in the sense of research) or research activities that have been undertaken with this collection. This remains the exclusive domain of the researcher, who needs to form her own overview of the research relevance of the collections.

There is often a lack of organisation of the available sources for the Holocaust and, if a collection has been organised, it will usually be done according to traditional archival principles. This implies that there are no references to related collections, related research projects or related publications. Furthermore, relevant material can be hidden in collections which have not been made accessible. A researcher will have to rely on her network to be alerted to this body of source material. Thus, for a researcher a good relationship with archivists as gatekeepers to the collections is often decisive for successful investigations.[7]

To support archival research better, EHRI will produce an integrated resource to hold Holocaust collection descriptions together with information relevant to research. So, for instance, while a section entitled 'scope and content' will summarize the general character of the holdings, a separate section will summarize the main Holocaust-related areas which are covered in the collection. All metadata produced are in accordance with ICA's ISAD(G) and ISDIAH standards[8], but enhanced with research-relevant information. The aim is to provide researchers with enough information for them to decide which institutions they need to visit to support their research. Before we discuss the information integration work further in the next section, we would like to briefly point out a more

immediate benefit from having an integrated resource dedicated to Holocaust research and how we support direct access to Holocaust material with travel grants.

The EHRI identification work aims to finally improve access to existing archival infrastructures in Holocaust research. An immediate benefit from a systematic investigation of Holocaust research material is to improve the physical access to these collections. A dedicated scholarship programme is a key activity to support and gain support from scholars. Researchers can apply through EHRI for a short-term fellowship currently at five research infrastructures: Institut für Zeitgeschichte in Munich, Jewish Museum Prague, Mémorial de la Shoah in Paris, NIOD Institute for War, Holocaust and Genocide Studies in Amsterdam and Yad Vashem in Jerusalem. These are all very established research infrastructures for Holocaust research.

Our experience is that in particular junior academics, Postdocs, PhD candidates with limited resources and researchers from Eastern Europe seem to take these trans-national access services, as they are called in the FP7 Integrating Activities programme. This part of EHRI provides direct opportunities for researchers. Even short-term travel grants can do a lot of good in Holocaust research, where due to the dispersion of Holocaust material researchers have to travel and visit several institutions. These short-term fellowships are considered a welcome complement to existing fellowships programmes offered by Yad Vashem and USHMM. Their programmes are more long-term and allow researchers to visit their institutions for respectively 2 to 4 and 3 to 9 months.

The EHRI call for applicants generated large interest from the Holocaust research communities. For the 2012 entry to the fellowship programme, overall 75 applicants from 22 countries have submitted a proposal. An international panel of experts has then selected 12 researchers on the basis of the excellence of their research proposals for a fellowship during 2012 at the five archival and research institutions mentioned above. The fellowships allow researchers to spend 4 to 8 weeks in one of these infrastructures. They research the collections and participate in scholarly exchanges with the staff of these institutions and beyond.

Linda Margittai was one of these fellows and chose NIOD as her destination in 2012. She applied for a fellowship to complement her research project, which examines the fate of Hungarian Jews living abroad during the Holocaust. Her stay at NIOD allowed her to work on her first case study to clarify which factors formed the fate of Jews of Hungarian citizenship living in the Nazi-occupied Netherlands. Reflecting on her experiences,[9] she said that the collections of NIOD, both archival and library collections, have provided her with very valuable materials, which have helped her ‘to clear up some essential research problems’. She was also very appreciative of the new academic contacts she had made. All in all, her stay at NIOD contributed to the development of her project and her academic network.

In the future, the EHRI fellows will also form a group of first users of the EHRI portal and research environment and contribute to the expansion of the virtual integration of Holocaust material, where the identification of Holocaust archives and collections is the first step to bringing these together into an integrated information resource The next section describes the infrastructure we provide for such a resource, and why we believe new graph databases have a lot to offer here.

Towards a new kind of infrastructure: the archive graph

Our work on identifying Holocaust archives and collections and use cases such as Terezín have taught us to understand that differences in practices and traditions in the Holocaust institutions will not disappear any time soon. Our aim is not to attempt to create a perfect world and to demand a close integration of all participants in Holocaust Research. Differences will always remain, as all the partners in EHRI are independent institutions with their own priorities and necessities linked to them.

Originally, we did not find it difficult to convince partners of the basic idea that Holocaust research material needs to be integrated and that this integration will enable better research. Examples such as the quoted case of Terezín are widely accepted and point to exactly such a necessity. More difficult has already been to emphasize the difference between an archive-driven interest in integrating Holocaust material and a research-driven one, as described above. Convincing institutions, however, to change their underlying archival integration systems and their approach to publishing their catalogues and finding aids online has proven to be almost impossible. This is why we had to come up with a different concept of a research-driven archival integration and revise our original thinking.

Once the relevant archives and their holdings are collected, the next step in the textbook approach for an integration of archival documents into a central resource would entail setting common standards for all partners and then somehow enforcing the implementation of these standards in situ. It has soon become clear in EHRI that this will not be possible for various reasons. A key one is that institutions differ heavily in the metadata they use and in the way they apply the metadata towards the various parts of their systems. A recently finished internal report of EHRI on existing metadata standards found that institutions used either the archival ISAD(G) standard or the more library-oriented Dublin Core to describe their collections. Some even use the archival encoding standard EAD[10] to exchange information between these institutions. This is the good news. The bad news was that half of the archives that responded to the survey said that they do not use archival standards at all. Among these were also some of the major archives in Holocaust research. It is unrealistic to assume that one can change the institutional approach, as this would imply a major cost for the institutions. We were therefore looking for a new integration approach that would require the least possible effort from the partners.

An integrating activity such as EHRI needs to confirm its partners' independence rather than negate it. EHRI wanted to investigate completely different paths towards this challenge. While we remained committed towards existing techniques from the open archives world such as OAI-PMH in combination with EAD[11] we started to investigate a new kind of infrastructure that would allow us to focus on the content as a researcher would need it, rather than on often abstract debates about metadata standards. We are committed towards taking the data as we found it, heterogeneous and often incomplete. It was important to us not to change the archival landscape in Holocaust research but to provide useful information for researchers. To this end, we started off from the basics and looked at what the descriptions essentially are beyond all the metadata standards that were incorporated in them. Again, it was the format of the FP7 Integrating Activities programme which allowed us to go down this route, as it encourages the creation of new ways of integrating resources.

Archival descriptions are essentially little pieces of texts enriched with varying amounts of complementary information about them such as the locations of the archives or titles for the descriptions. These texts were connected with each other by their archival context. This means they are structurally connected by belonging to collections, to repositories, etc. In the archives', however, world they are seldom connected by their content, research activities, etc. These are often not recorded in archives and are the first contribution to research EHRI will make, as described above.

As we had to discard our original approach of integrating these materials, we looked for alternatives that would allow us to integrate these kinds of small texts and found them in the world of social web technologies and advanced text technologies such as OCR and text analysis. For the remainder of this article we will analyse the current state of our work and experiments in these two areas. The final part then describes our plans for a larger virtual research environment constructed from these experiments.

Finding aids are just one example of a typical type of research material in the humanities. Palmer et al.[12] have investigated the types of information source materials used in different humanities disciplines, based on results contained in the US Research Libraries Group (RLG) reports. Structured data is relatively little used, and data as it is traditionally understood in the sciences, i.e. the results of measurements and the lowest level of abstraction for the generation of scientific knowledge, even less so. As we have described elsewhere in detail,[13] historical research relies not on measurements as a source of authority, but rather on the provenance of sources and assessment by peers.

Indeed, studies of the research behaviour of history scholars[14] have demonstrated that they continue to rely on primary materials held in dedicated collections in special places, namely in archives, and it is in archives that the scholar carries out the work of assessing these source materials. Archival records are primary sources about the past and may take many forms, such as government correspondence, financial documents, photographs, sound recordings, etc. All this information is unstructured in its nature and is accessible via finding aids, which themselves are not structured information, but are, in most cases, documents containing detailed information about the records in a specific archival collection.[15]

Many traditional archival management systems seem to neglect this essentially document-oriented nature of archival material, as they are based on standard database technologies.[16] As such, they not only assume a structure in the data but also need to plan ahead very carefully, because database systems are based on a static schema description of the underlying information. This has many advantages; for instance the ability to enforce consistency and ensure that transactions are safe. However, in the world of archival integration of non-standard heterogeneous material these are often secondary problems. The static schema that underlies a database can on the contrary often be identified as a main obstacle to the integration, as it enforces a God-like view of all sources beforehand, where everything needs to be made to measure according to this schema.

For Holocaust research, challenges such as these have led in the past to either the outright rejection of a computational integration approach as too complex or the reduction of the resulting integrated information to a small number of common fields one can assume to exist across all archival databases.[17] Both are not satisfactory for research. The first one is not satisfactory, because it neglects the essential benefit of integrating Holocaust material that is so evident from examples

such as Terezín. The second one is not satisfactory, because it implies a reduction of the information available to researchers to such a low level that it is essentially useless. We need systems that can grow harmoniously with every new piece of information we can gather and not reduce the amount of processesable information from the outset. We therefore looked at alternatives to existing database technologies and found them in so-called NoSQL databases, powering many of the current large-scale social media sites on the web.

NoSQL databases are a recent innovation[18] and often give up on the schema-centricism of traditional databases. They realise instead a more flexible model, which concentrates on describing heterogeneous information such as texts in archival descriptions.[19] More famous example implementations of NoSQL databases include Amazon's Dynamo store[20] that realises a simple key-value store, where anything can be a key to a stored data. Furthermore, there is Google's Bigtable approach[21], where data is stored in one big table, but each of the rows in the table can have its own schema. We have concentrated on a particular class of document-centric NoSQL databases called graph databases, as they fulfil our second requirement that new information and new relations can be added easily.

With graph databases, one is not restricted in the information one can attach to documents as it is the case for a traditional database environment, where each document is embedded into a static schema.[22] Within traditional databases, a document like a finding aid entry is simply another element in a row that otherwise contains a lot of other elements. Only some of these entries might be known about this particular document, which means that there are a lot of empty entries in the database. To avoid this situation, graph databases are based on the more flexible Euler's graph model with nodes and edges between them and all the standard well-established means of graph processing. The model[23] that underpins the graph database has three simple elements: nodes and their relationships as well as properties that can be given to both nodes and relationships[24]. Graph database are thus quite an old idea[25] but only since the advent of the document-centric databases and the social web, there is a renewed interest in developing them into full-scale storage infrastructures.

Graph databases add to other types of NoSQL and document-centric databases the ability to integrate several different types of documents relatively effortless. As one can see from our discussions, the NoSQL approach gained popularity when completely new problems of big data and its integration in the social web and on the web in general occurred. We have chosen graph databases instead of more traditional document-centric databases because they scale best towards complexity or towards information that is not uniform[26] and as 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[27], most humanities data possesses a complex structure, with many internal relationships both structural and semantic. Archival data for historical research is no exception and is highly contextual, its interpretation depending on relationships to other resources and collections (not necessarily digital). This internal connectedness also means that it is difficult to scale the processing necessary to analyse this data, as it cannot be easily distributed across different machines. But, we have found that in the humanities we do not need to care much about these problems, as in 80% of the cases

we do not need to use the kind of processing that would require us to think about parallel processing requirements.[28] Many humanities resources are create-once-read-many, where enhanced processing is mainly needed to create and/or enhance resources and can therefore be realised as resource services, which are currently external to the EHRI requirements.[29]

Implementation of the research graph

The choice to implement the EHRI integrated information resource with graph databases was therefore easy, as it comes with a rich data model that makes modelling a domain much simpler than in other traditional database approaches. It is easy for researchers to imagine the set of documents on the Holocaust as gigantic graph of nodes that are connected via various relationships. The important question here is not that the nodes are connected but how. So, for instance, one would like to describe that document X is connected to document Y, as they are about prisoners in the same camp, while they might be distributed across various archives.

With graph databases researchers can therefore be involved directly in modelling their domain without the need to learn new techniques such as entity relationship models, etc. Archival documents are the nodes in the graph database, while the edges are the flexible relationships between them. However, graph databases pose the other challenge that little work has been done yet with them in the field of research computing for history. This means we basically need to develop our own system from scratch, which makes it a higher development risk. Particularly challenging is transaction management, which is a big issue in the world of integrating heterogeneous data sets and integrating data sets from several remote sources.[30] So, from an implementation point of view graph databases are definitely not the easy way. But, they make much easier the task we experienced as the most challenging one in our field, which is to translate the model in a research domain in the head of a researcher into something a computer can process. Here, the network abstraction behind graph databases seems to be much closer to how a humanities researcher sees her own field.

Similarly, researchers can retrieve much more intuitively the data out of the document store. In another experiment at King's College London, evaluation by researchers found that graph databases were taken up keenly by researchers in Classics[31], as key concepts and documents of the classical world in this case could be browsed easily and contextualised with other related concepts. For instance, a researcher might have been recording ancient inscriptions on-site and wish to research or contextualize particular words in them. They might wish to identify inscriptions, which refer to a particular person (e.g. a particular Roman Emperor), and filter these by certain attributes, such as time period or location. All this is much easier with a graph model rather than a simple gazetteer or other reference resource. This positive experience has encouraged us to continue with the development of a graph database infrastructure for EHRI. As seen in our experiments beyond the more technical view of the underlying integration infrastructure, graph databases scale towards the kind of queries research would like to ask better than traditional databases.[32]

The Terezín example shows the interconnectedness of Holocaust research material. Even, if this interconnectedness is currently not well represented in archival records, many Holocaust research

questions lend themselves to network topologies, which is where for us the new research questions and future research lies. In traditional archival databases, it is difficult to scale questions such as all administrations involved in the arrest of camp prisoner X or all documents related to administrations in a particular place Y. This is because links in traditional relational databases are only indirectly represented using link tables and almost all linking in such a database requires to run full-scale relational joins, which are computationally expensive. Graph databases treat all entries as part of a large-scale graph and provide efficient graph algorithms to traverse these entries.

Relationships are first class citizens in graph databases. They can be typed and one can even assign a key to the relationship. This enables effective querying of the contextual relationships for Holocaust research interests. Queries like 'find all Dutch prisoners in Terezín which came from Amsterdam and were first imprisoned in 1941' might under normal circumstances require multiple deep joins in relational databases across many tables. This makes them perform very badly, while for graph databases efficient graph algorithms can traverse the data universe (modelled as a graph) quickly and it is easy to represent structures such as hierarchies, which dominate the world of archives but are difficult to fit into the relational model. Graph database provide an excellent underpinning of the traditional browsing tasks in historical research.

Next to browsing a graph, graph databases also offer technologies for more traditional searching, as it is another standard requirement for contemporary historical research.[33] Graph databases work well with some advanced search techniques such as facetted searching and filtering.[34] Our graph database tightly integrates with SOLR.[35] Any nodes and their properties can be indexed, which next to generic text-based searches can lead to very effective browsing possibilities for the underlying document space. Subgraphs can be flexibly combined by using external indexes. In related work at US Holocaust Memorial and Museum (USHMM), SOLR was used to integrate all archival information systems into a combined search interface. We work closely with USHMM to optimise the integration of SOLR with various content and metadata standards in the field of Holocaust research.

With SOLR, we can support better more advanced means of access to facts in the documents and enable deep semantically meaningful access to the documents. Semantically enriched library and archive federations have recently become an important part of research in digital libraries and archives.[36] Especially so, as research users often have more demands on semantics than is generally provided by archival metadata. For instance, in archival finding aids place names are often only mentioned in free-form narrative text and not especially indicated in controlled access points for places. Researchers would like to search for these locations.

Extraction of place names from the archival descriptions might support this.[37] Place names would also integrate well in a graph database as they provide first-class relationships between documents and can be added flexibly to nodes and relationship as further information about them. For the future EHRI infrastructure, we want to use information extraction services to enrich the researchers' experience.[38] In our experiments, we concentrated on extracting names and places facets, both immensely important for Holocaust research. As most of the relevant finding aids, however, are still in printed form, we investigated furthermore the use of an open source OCR infrastructure developed at King's College London for the Ocropodium project.[39] Current commercial OCR technology does not serve well specific research interests in historical document collections, as it cannot be easily customised.

Extracting semantic information from low quality textual data is challenging, which is why we plan to set up our own open source OCR infrastructure.[40] Most commercial OCR software products are proprietary 'black boxes' which provide digitisation staff with little scope for understanding their behaviour and customising parameters under which they run. We had hoped that with open source tools the information extraction process could be improved. Our initial evaluation results[41] show that even within a standard setup and without further training and customisation of the tools, we can extract useful information. With a few added components such as the integration of advanced gazetteers of place and person names of the Holocaust we could easily improve the initial results further. But, ideas like the OCR infrastructures are currently just plans for EHRI based on past work, which we would like to follow more closely once we have gained a better understanding of the underlying data.

Conclusions

In this article we presented our ideas on an integrating activity for Holocaust research on archives. We offered details on the European Holocaust Research Infrastructure project that attempts to overcome both physical as well as digital barriers to access Holocaust research collections. The starting point of our investigations was that in the current situation Holocaust researchers are not able to make the necessary links between collections. Research material is dispersed across different countries and archives because of the attempts by the Nazis to destroy the evidences and because the war has led to large scale displacements and migrations movements. The Adler documentation on Terezín was cited as a typical case of the sometimes surprising links and connections between different archival locations as well as the partly chaotic way in which collections were established in early documentation efforts.

Historical collections remain hidden in two ways. Firstly, they are simply not yet recorded. Even if they are recorded, their context and therefore their meaning will stay unknown. EHRI is to our knowledge the first systematic attempt to mitigate this by intensifying the identification effort and by providing new and innovative digital means to link these collections. We presented our more traditional efforts such as the fellowship programme but also the digital research infrastructure we have set up to link collections. We have followed new ideas here coming from the world of integrating social media activities and found the model to work well in the world of heterogeneous collection descriptions of Holocaust research. To us, the innovation of EHRI lies in the combination of digital and non-digital means to integrate existing infrastructures. We believe this might be a model for many related research activities in the humanities.

For the next steps after the identification and basic collection integration work, we have run various experiments that will allow us to enable a strong document-centric approach in Holocaust research. We are confident that we have managed to set up the necessary infrastructure but at the moment our work has been limited to collection level descriptions. Next, we aim for more direct work with documents. We do not believe that collection descriptions will disappear. They themselves provide essential context to Holocaust researchers, and the idea of a completely digital documentation landscape for research remains a dream for the foreseeable future. But historians expect nowadays access not just to collections but to documentation where it is digitally available. The traditional distinction between collection-level and document-level documentation is disappearing fast in a

digital environment. So, our next steps will be to become an integrating activity not just across archives but within archives themselves and link sources in whatever form they appear.

References

[1] For a discussion of the particularities of a humanities research infrastructure see the contribution by our EHRI colleagues Reto Speck and Petra Links elsewhere in this special issue.
[2] W. Duff, B. Craig and J. Cherry, 'Historians' use of archival sources: promises and pitfalls of the digital age', *The Public Historian*, 26, no. 2 (2004), 7-22.
[3] An extended discussion of the dispersion of records of Terezín will be published in Heike Neuroth, Norbert Lossau, Andrea Rapp (eds.), *Evolution der Informationsinfrastruktur: Forschung und Entwicklung als Kooperation von Bibliothek und Fachwissenschaft* (Göttingen, in preparation).
[4] H.G. Adler, *Theresienstadt, 1941-1945: Das Antlitz einer Zwangsgemeinschaft. Geschichte, Soziologie, Psychologie* (Tübingen, 1955).
[5] Franz Hocheneder, *H.G. Adler (1910-1988). Privatgeleghrter und freier Schriftsteller* (Vienna, 2009).
[6] Within EHRI, we are currently working on a comparative research project on finding aids in different archives. The results are planned for publication in the international journal *Archival Science*.
[7] Duff, Craig and Cherry, 'Historians' use of archival sources', 7-22.
[8] J. Dryden, 'Two new ICA descriptive standards: ISDF and ISDIAH', *Journal of Archival Organization*, 7, no. 3 (2009), 129-32.
[9] European Holocaust Research Infrastructure, *Newsletter*, http://www.ehri-project.eu/ehri-newsletters/october-2012, last accessed 7/11/2012.
[10] K. Kiesling, 'EAD as an archival descriptive standard', *American Archivist*, 60, no. 3 (1997), 344-54.
[11] C.J. Prom and T.G. Habing, 'Using the open archives initiative protocols with EAD', unpublished paper presented at the Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries, 2002.
[12] C.L. Palmer, L.C. Teffeau, and CM Pirmann. "Scholarly Information Practices in the Online Environment: Themes from the Literature and Implications for Library Service Development. Report Commissioned by Oclc Research (Dublin, Oh: Oclc, 2009)." available online at www.oclc.org/research/publications/library/2009/2009-02.pdf, last accessed 20/3/2010.
[13] T. Blanke, L. Candela, M. Hedges, M. Priddy, and F. Simeoni, 'Deploying general-purpose virtual research environments for humanities research', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 368, no. 1925 (2010), 3813-28.
[14] Duff, Craig and Cherry, 'Historians' use of archival sources', 7-22.
[15] D.V. Pitti, 'Encoded archival description: The development of an encoding standard for archival finding aids', *American Archivist*, 60, no. 3 (1997), 268-83.
[16] J.A. Levine, J. Evans and A. Kumar, 'Taming the "Beast": An archival management system based on EAD', *Journal of Archival Organization*, 4, no. 3-4 (2007), 63-98.
[17] M. Vernooy-Gerritsen, *Emerging standards for enhanced publications and repository technology: survey on technology* (Amsterdam, 2009).
[18] M. Stonebraker, 'SQL databases v. NoSQL databases', *Communications of the ACM*, 53, no. 4 (2010), 10-11.
[19] N. Leavitt, 'Will NoSQL databases live up to their promise?', *Computer*, 43, no. 2 (2010), 12-14.
[20] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall and W. Vogels, 'Dynamo: amazon's highly available key-value store', unpublished paper presented at the ACM SIGOPS Operating Systems Review, 2007.
[21] F. Chang, J. Dean, S. Ghemawat, W.C. Hsieh, D.A. Wallach, M. Burrows, T. Chandra, A. Fikes and R.E. Gruber, 'Bigtable: A distributed storage system for structured data', *ACM Transactions on Computer Systems (TOCS)*, 26, no. 2 (2008) 4.
[22] R. Angles and C. Gutierrez, 'Survey of graph database models', *Computing Surveys*, 40, no. 1 (2008) 1.
[23] Angles and Gutierrez, 'Survey of graph database models', 1.
[24] J. Webber, 'A programmatic introduction to Neo4j', unpublished paper presented at the Proceedings of the 3rd annual conference on Systems, programming, and applications: software for humanity, 2012.
[25] Angles and Gutierrez, 'Survey of graph database models', 1.
[26] E. Eifrem, 'Neo4j—the benefits of graph databases', *no: sql (east)* (2009).
[27] T. Blanke, L. Candela, M. Hedges, M. Priddy, and F. Simeoni, 'Deploying general-purpose virtual research environments for humanities research', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 368, no. 1925 (2010), 3813-28.

[28] T. Blanke and M. Hedges, 'Humanities e-Science: From systematic investigations to institutional infrastructures' (paper presented at the e-Science (e-Science), 2010 IEEE Sixth International Conference on, 2010).
[29] T. Blanke, M. Hedges and R. Palmer, 'Restful services for e-humanities', unpublished paper presented at the 3rd IEEE International Conference on Digital Ecosystems, 2009.
[30] S. Venugopal, R. Buyya and K. Ramamohanarao, 'A taxonomy of data grids for distributed data sharing, management, and processing', *ACM Computing Surveys (CSUR)*, 38, no. 1 (2006) 3.
[31] T. Blanke, M. Bryant, S. Dunn, G. Bodard, M. Jackson, and D. Scott, 'Linked Data for Humanities Research --- The SPQR experiment' (paper presented at the IEEE International Conference on Digital Ecosystems and Technologies for Complex Systems, Campione, 2012). .
[32] Eifrem, 'Neo4j—the benefits of graph databases'.
[33] T. Blanke, L. Candela, M. Hedges, M. Priddy, and F. Simeoni, 'Deploying general-purpose virtual research environments for humanities research', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 368, no. 1925 (2010), 3813-28.
[34] Eifrem, 'Neo4j—the benefits of graph databases'.
[35] Webber, 'A programmatic introduction to Neo4j'.
[36] S.R. Kruk and B. McDaniel, *Semantic Digital Libraries* (Berlin, 2009).
[37] Blanke, T., M. Bryant, and M. Hedges. 'Open Source Optical Character Recognition for Historical Research' Journal of Documentation 68, no. 5 (2012): 659-683.
[38] K.J. Rodriquez, M. Bryant, T. Blanke and M. Luszczynska, 'Comparison of named entity recognition tools for raw OCR text', Proceedings of KONVENS 2012 (LThist 2012 workshop), (Vienna, 2012): 410-414.
[39] M. Bryant, T. Blanke, M. Hedges and R. Palmer, 'Open source historical OCR: The OCRopodium project research and advanced technology for digital libraries', in M. Lalmas, J. Jose, A. Rauber, F. Sebastiani and I. Frommholz, ed., *Lecture notes in computer science* (Berlin / Heidelberg, 2010).
[40] T.L. Packer, J.F. Lutes, A.P. Stewart, D.W. Embley, E.K. Ringger, K.D. Seppi and L.S. Jensen, 'Extracting person names from diverse and noisy OCR text', *Proceedings of the fourth workshop on Analytics for noisy unstructured text data* (Toronto, 2010).
[41] Rodriquez, Bryant, Blanke and Luszczynska, 'Comparison of named entity recognition tools'.