European Holocaust Research Infrastructure
H2020-INFRAIA-2014-2015
GA no. 654164

Deliverable 14.1

Report on research use cases

Giles Bennett
Institute for Contemporary History (IfZ)

Tobias Blanke
King's College London

Mike Bryant
King's College London

Martijn Eickhoff
NIOD Institute for War, Holocaust and Genocide Studies

Conny Kristel
NIOD Institute for War, Holocaust and Genocide Studies

Daan de Leeuw
NIOD Institute for War, Holocaust and Genocide Studies

Reto Speck
King's College London

Veerle Vanden Daelen
CEGESOMA

Start: June 2015
Due: April 2016
Actual: May 2016

EHRI is funded by the European Union

Document Information

Project URL	www.ehri-project.eu
Document URL	[www……]
Deliverable	D14.1 Report on research use cases
Work Package	WP14
Lead Beneficiary	P1 - NIOD-KNAW
Relevant Milestones	MS1
Dissemination level	Public
Contact Person	Martijn Eickhoff, m.eickhoff@niod.knaw.nl, +31(0)20-5233800
Abstract (for dissemination)	This deliverable presents six research use cases that have been defined based on the current state of the Holocaust historiography and Holocaust-related digital sources available through EHRI. Besides historical research these research use cases also include the other two disciplines in EHRI: archival science and digital humanities. These six cases will inform the method and tool development work undertaken in tasks two and three of the work package. This deliverable also takes into account the size and type of relevant archival collections and the legal and ethical requirements in working with these documents. In sum, the research use cases aim to facilitate the co-investigation and co-development of digital methods and tools between Holocaust researchers and digital specialists.
Management Summary	(required if the deliverable exceeds more than 25 pages) [Max. 500 words]

Table of Contents

1 Introduction

Work Package (WP) 14 will provide services for conducting research with digital archival sources (a digital historiography) that address major challenges faced by Holocaust historians: the amount of potentially relevant source material to be considered; the semantically complex and dispersed nature of such sources; and challenges arising from collaborative approaches to Holocaust research. This deliverable is the result of the first of three complementary tasks within WP 14 looking for the means to develop digital historiographies of the Holocaust. In this deliverable, six research use cases based on the current state of the historiography and Holocaust-related digital sources available through EHRI have been defined. These six use cases will inform the method and tool development work undertaken in tasks two and three but they are also independent from these tasks as they rely on further digitisation work to be undertaken.

1.1 Research questions/problems to be addressed

The starting point of the research use cases developed as part of WP 14 are current debates and pending research questions in Holocaust historiography. Holocaust studies increasingly develop into a broad field of interdisciplinary research in which a wide range of topics are addressed. Relevant current developments - as identified in the recent handbooks of Holocaust researchers Dan Stone, and Frank Bajohr and Andrea Löw[1] - are: 1) the relation between central and local decision making; 2) a focus on local histories or histories from below; 3) a focus on agency and networks of victims; 4) the spatial turn which focuses on the role of spatial structures in the organisation and the post war impact of the Holocaust, with a special focus on the related migration movements of refugees and displaced persons; 5) the role of bystanders; 6) the forensic approach, which focuses on killing sites and corpses of mass violence and genocide; 7) the cultural turn, which questions which meaning was given to the Holocaust by perpetrators/ bystanders/ victims; 8) the relation between the Holocaust and colonialism; 9) the role of ideology in the Holocaust; 10) material aspects of the Holocaust related to the confiscation and restitution of property; 11) the relation of the Holocaust with civil war, armed resistance and violence against non-Jewish people.

These active areas can be enhanced by digital methods as demonstrated in related fields:

- Social Network Analysis has helped businesses understand the relation between central and local decision-making. It has also been successfully applied in the social sciences to understand agency and networks; who might be classified a bystander to particular activities or how has armed resistance been formed. Social Network Analysis has finally been used to understand the links between elites in colonialism and imperialism.
- Text analysis has its origins in computational linguistics in order to work with large collections of textual materials. It has been successfully applied studying racist language and sentiments in ideologies and discourses. By unlocking large textual collections, text analytics is furthermore probably the only way to enable access to the large unknown collections, that historians from below and local historians are interested in. It has finally proven to be an excellent tool in cultural studies for a new perspective on how meaning is given in scientific and other discourses to particular concepts such as perpetrators and victims.

[1] Dan Stone (ed.), *The Holocaust and historical methodology.* Making Sense of History. Studies in Historical Cultures (New York / Oxford 2010); F. Bajohr and A. Löw (eds.), *Der Holocaust. Ergebnisse und neue Fragen der Forschung* (Frankfurt am Main 2015).

- The spatial turn finally has only been made possible by GIS technologies that allow to answer questions about whether spatial features such as locations or land cover can help us understand social and historical events in many areas of social sciences. These technologies are also key to decoding the forensics of places where victims were murdered, buried, etc. as well as other material aspects. They have thus become part of the arsenal of archaeological excavations.

While all the eleven areas described above are thus of interest to a digital historiography, the first five developments are particularly important topics for Holocaust scholars because they integrate the central themes of Holocaust Studies, which are the actions of perpetrators, the fate of the victims and the behaviour of bystanders. At least four out of the six research use cases in this deliverable can be connected to a number of these developments: the first topic on the relation between central and local decision making relates to the research use case 'Between Decision Making and Improvisation. Tracing and Explaining Patterns of Prisoners' Transfers through the Concentration Camp System'; the second topic of local history is integrated in all historical research use cases; the third topic of the agency of victims relates to the 'Names and Networks' research use case on Jewish survival chances; and the fourth topic on the spatial turn informs the research use cases 'People on the Move' and 'In Search of a Safe Haven'. In almost all cases, the perspectives and experiences of prosecuted Jews are of pivotal importance.

The WP agreed at the beginning of its work on the following principles for the research use cases:

- They should mainly cover historical research but also include the other two disciplines in EHRI: archival science and digital humanities.
- Rather than limiting us to existing digitized material, the research use cases should imagine what would be possible in digital historiography if particular materials were fully digitised. Furthermore, simple digitisation is often not enough, the material has to be available in formats that can be actioned by computers. Many of the research use cases therefore also address the creation of new derived data sets for these purposes.
- Most importantly, all research cases should be guided by current research questions in the three disciplines in EHRI.

All research use cases were gathered from researchers within the EHRI consortium. They were provided with a template and asked to limit themselves to research that is empirical, i.e. that is based on investigating existing collections. The submissions were then centrally collected by NIOD and reviewed by the WP, supported by the EHRI Project Management Board and Executive Team.

1.2 Size and type of relevant collections

The research use cases presented in this deliverable all take substantial archival collections as a starting point. In most cases the data from these collections will be connected across more than one archive. They all also demonstrate that substantial further efforts are often needed to prepare the data in the archival collections before digital methods can be applied.

The research use case 'Names and Networks. Chances of Survival during the Holocaust' builds on the digitized Jewish Council Archives in Europe project and is carried out in close cooperation with EHRI, Digital Monument to the Jewish Community the Netherlands (Digitaal Monument Joodse Gemeenschap Nederland), and the United States Holocaust Memorial Museum (USHMM) which administers the list of Dutch Jewish survivors. The research use

case ‘In Search of a Better Life and a Safe Haven: Tracing the Paths of Jewish Refugees (1933-1945)’ builds on various archives of inhabitant records of local municipalities. The research use case ‘People on the Move. Revisiting Events and Narratives of the European Refugee Crisis (1930s-1950s)’ will use a digitized sample of data from at least 10.000 persons from the collection of the International Tracing Service (ITS). The research use case ‘Between Decision Making and Improvisation. Tracing and Explaining Patterns of Prisoners’ Transfers through the Concentration Camp System’ starts with a digitized reconstruction of the WVHA-Häftlingskartei consisting of 148.782 record cards kept at the KZ-Gedenkstätte Flossenbürg. The digital images of these prisoner cards were made in: Bundesarchiv (Berlin), Polish Red Cross (Warsaw), State Museum Auschwitz-Birkenau (Oświęcim), State Museum Stutthof, (Sztutowo). The research use cases Archives and Machine Learning as well Networked Reading start from archival and digital humanities research perspectives and will therefore take a slightly different approach. For ‘Archives and Machine Learning’, we will mainly rely on the existing EHRI repository of collection descriptions, while Networked Reading will lead to a toolset to analyse more than one digital collection of textual material. Here, we have already begun to investigate memory processes in survivor testimonials and their internal and external linkage. The use case will mainly investigate how to transform such material into reliable data sets for historical research.

1.3 Legal and ethical requirements

The research use cases build upon a varied collection of digitized Holocaust related archives kept in and beyond Europe. Potential research teams will study and respect the legal conditions of the different ‘hosting’ countries and institutes with regard to the archives in general and data in particular. During the research the digitized data – in particular those that include personal information – will be handled with the greatest care. Therefore the research material will only be accessible to members of the project team. Sometimes additional regulations are to be taken into account. In case of NIOD collection 182 (Joodse Raad voor Amsterdam), which is a crucial digitized archive for the research use case ‘Names and Networks. Chances of Survival during the Holocaust’ we are, for example, able to include sensitive personal files in our research, as long as the research takes place within the firewalls of NIOD. Every research team will keep a log that describes the way the members deal with legal and ethical requirements during the different project phases . This is done in order to develop a more general best practice protocol. Registering such ethical and legal requirements for each use case, will be an essential part of the work on all use cases.

2 Names and Networks. Chances of Survival during the Holocaust

2.1 Purpose

This project aims to investigate the networks in which European Jews operated during their persecution in the Second World War. Analyzing these networks makes it possible to improve our understanding of the various chances of survival that persecuted Jews all over Europe had and the kinds of dislocations, deportations and migrations that occurred before and during the Holocaust. Most persons needed the support of other people to survive. Persons found aiders inside their personal and group networks.

When a scholar wants to analyze group behaviour, it is very challenging to investigate individual motives of historical actors belonging to a certain group. However, through an analysis of the networks of these historical actors it becomes possible to develop a deeper understanding of their motives and collective behaviour as members of a group. Network analysis can be defined in this case as an investigation of a historical group linked by a common factor based on the connections between individual members of this group. The crucial question is the way these members operated within and upon the social, political, legal, economic, and intellectual institutions of their time. Instead of a limited focus on a small number of individual biographies, the network analysis approach focuses on patterns of activities and interrelationships within a historical group. With the application of digital tools and techniques, large amounts of source material can be integrated. This project aims to develop methods for writing collective biographies of the victims of the Nazi persecution. In this project, we will start with Jews from Amsterdam. This is of high relevance because in the Netherlands an extraordinary number of Jews (75% of the pre-war population) were murdered during the Holocaust. Additionally we aim at including different cases from other occupied countries. Our approach enables us to get beyond the stereotyping of national reactions toward the persecutions of Jews; especially because of its strong local and social dimension.

The following research questions can be addressed: To what extent did Jews participate in various types of networks? Does an analysis of the networks indicate that there is a relation between taking part in networks and survival? What kinds of networks did religious and secular Jews have and how did this influence survival? Were some networks more important for survival than others? Were political networks more important than social or economic ones, or vice versa? Was participation in non-Jewish networks particularly helpful?

2.2 Materials and sources

The project will use a variety of data. The starting point is the NIOD project Jewish Council Archives in Europe. This project brings together important – and in some cases still unexploited – archival collections on Jewish Councils and associations from Belgium, Czech Republic, Finland, Hungary, Netherlands (Amsterdam and Friesland) and Poland. The main objective is to prepare these collections for use by researchers and the general public, now and in the future. Focus will be on digitization of original documents and creation or improvement of digital finding aids. By reformatting the contents from analogue to digital carriers, the digital master files will serve as surrogates for future use. The Jewish Council Archives in Europe project is carried out in close cooperation with EHRI (please, compare: http://www.niod.knaw.nl/nl/projecten/jewish-councils-europe).

Furthermore, two more collections can be incorporated into this research:

1. The dataset of the Digital Monument to the Jewish Community the Netherlands (*Digitaal Monument Joodse Gemeenschap Nederland*), in which extensive biographical information on individual persecuted Dutch Jews has been collected.[2]
2. The records of persecuted Jews held at the EHRI partner ITS (International Tracing Service). They can likewise provide information on the networks in which Dutch (and other) Jews participated.
3. The list of Dutch Jewish survivors held at our partner institute the United States Holocaust Memorial Museum (USHMM) will enable us to identify survivors and analyse the networks in which they participated. See: http://www.ushmm.org/online/hsv/source_view.php?SourceId=27995 and http://www.jewishgen.org/databases/Holocaust/0239_Dutch_survivors_lists.html.

It is expected that the majority of the documents written in German and Dutch, which are both covered well by existing language technologies. At the beginning of the project we will make a sample of the documents to check this assumption.

2.3 Procedure

In order to make an outline of the steps we will follow to achieve our research goals, we start making two observations about the current state-of-the-art in digital network analysis for history:

1. Network information in history is currently commonly modelled using a database that represents *assertions* about individuals identified in the texts. While some of these assertions are represented by dedicated columns within the database schema – assertions that are common to all people, such as birth, death, kinship or occupation, or are considered particularly relevant for the prosopography being developed – in most cases the assertions consist of free narrative text, for example to the effect that X did something, or X did something to Y. This has at least two consequences: on the one hand, the use of fixed schemas makes such resources difficult to extend or combine. On the other, variability and lack of agreed meaning that arises from not using formal vocabularies make such resources more ambiguous and difficult to use.
2. Network analysis may involve a labour-intensive, manual process of reading through source material, extracting references to individuals and assertions about them, and entering them into the database. While this may be effective when dealing with relatively small corpora, for example in ancient contexts where less material has survived, for more recent periods in history the rapid increase in the amount of digital material available – whether the results of large-scale digitisation programmes or born-digital documents such as emails – renders such approaches impractical. However, this is an opportunity as much as a challenge.

The novelty of our approach will be the development of network analysis methods that (a) use novel representations to facilitate the analysis of social and other networks, and (b) use computational techniques (Natural Language Processing) to deal with very large quantities of source material. To address (1), we will adopt a “Linked Data” information model that underlies current thinking about the Semantic Web. The key idea behind Linked Data is that knowledge can be represented as a tagged network or graph structure, whose nodes are named (or anonymous) entities corresponding to people, places, events and other objects of interest, and the arcs between the nodes are tagged with terms from ontologies (formal vocabularies) that correspond roughly speaking to the “verbs” in the assertions.[3] Linked data thus addresses the problem identified in (1) by formalising assertions as relationships

[2] See: http://www.joodsmonument.nl/?lang=en.
[3] See http://www.digitalclassicist.org/wip/wip2010-03th.pdf for a similar approach (for ancient history).

between entities, although this can result in significant complexity when we later seek to interact with the resulting knowledge structure. However, the development of tools supporting such interactions makes such an approach a good trade-off, particularly for projects such as ours that seek to integrate knowledge from a broad range of disparate resources.

To address (2), a toolkit will be required for the creation and enrichment of network resources, integrating text mining tools and services to automatically tag and disambiguate the mentions of known entities, as well as discover new entities that need to be added to the knowledge base (e.g. a new person or the name of their mother/father, who are not known already), and to identify relationships between these entities.

2.4 Expected observations

First of all, like stated above, we hope to answer the questions whether, how and when specific networks contributed to the survival of Jews. Our assumption is that networks generally played a crucial role in survival rates since there is ample evidence that people needed other people aid them in the avoidance of being arrested and ultimately transported to concentration camps. Furthermore, we hope to specify which networks provided support and which were less helpful for surviving, and how the value of these networks changed over time and space. In the Amsterdam case it is expected that Jewish networks were more important for survival than non-Jewish networks, because most of the non-Jewish inhabitants of the Netherlands were rather indifferent to the National Socialist persecution of Jews. However, it is also possible that Jews who had particular good relationships with non-Jewish citizens could use these networks to contribute to their survival. In sum, after this research we hope to better understand the various survival modes and chances Jews had during their persecution in different parts of Europe.

2.5 Results / Analysis

We will produce the following research outputs:

- A new approach to representing network-related information related to the Holocaust
- A new methodology for reconstructing networks in the Holocaust that uses text mining and natural language processing across very large corpora of archival texts.
- A demonstration how network analysis can aid contemporary Holocaust research

3 In Search of a Better Life and a Safe Haven: Tracing the Paths of Jewish Refugees (1933-1945)

3.1 Purpose

During the Second World War, many Jews found themselves in places other than the ones they grew up in. For some, migration had started before the Nazis came to power, as individuals and families decided to move in order to build a better life, away from economic and/or anti-Semitic discrimination and violence. Others migrated after they came under Nazi-rule. During the Second World War, Jews were subject to forced displacements. The anti-Jewish measures under the Nazi-regime led to further attempts to flee and find a safe haven. As such, Jews who were born in Austrian Galicia, Lithuania or Germany may have first found refuge in the Netherlands or Belgium, attempted later on to reach Switzerland, but were eventually caught and deported from France to Poland. This research aims to map and better understand the different migration trajectories and to determine the factors that played a role in the migration movements of the migrants.

3.2 Materials and Sources

- Administration of citizens
- Registrations of Jewish communities
- Obligatory registrations during the Second World War
- Deportation lists
- Migration files (lists of entering foreigners etc.)

All of the archives in EHRI are potential use cases for this topic. Ideally, the corpus would consist of at least two different locations so that connections can be made between people equally documented in two related archives.

Not only the type of fonds named above, but also the descriptions of other collections can be of use for this topic. A quick look at the collection descriptions held at Kazerne Dossin (https://portal.ehri-project.eu/institutions/be-002157/search?page=1), for instance shows that these already include many names of people and places, which reveal migration histories. Also, the authority records/descriptions of the creators of collections (especially when dealing with private archives, where these authority records can contain biographical information on the creator of the collection which hint to his or her migration history) can be good step stones towards the sources named above.

3.3 Procedure

Information about the same person has been registered in multiple places; often in registration documents, which are written in different languages and even in different scripts. The challenge will be to structure and connect the data, which relate to the same person and his or her family. The more efficient one can look for one place or person across the data in different archival institution, the better the results for the analysis will be.

3.4 Expected observations

Via the procedure outlined above we hope to connect the sources on persons and places to answer research questions such as:

- When did migrations take place?

- To what degree can we identify chained migration (be it within a family or interpreted more largely as chained migration for the same place of departure)?
- To what degree are migration time, place of origin, number of family members (or people with the same migration history) in one place, and other factors potentially related to chances of survival during the Second World War?
- How many different language environments did the migrants/refugees encounter on their journey?

3.5 Results/Analysis

This research will outline paths of migration in times when more and more borders were closed off and possibilities were shrinking day by day. Which routes were followed by whom and by how many? Did the refugees indeed heighten their chances of survival by leaving their homes? Did their leaving have any effect on chances of survival for family members who stayed behind?

4 People on the Move. Revisiting Events and Narratives of the European Refugee Crisis (1930s-1950s)

4.1 Purpose

This project aims to investigate migration movements of European refugees from the 1930s until 1950s. The twentieth century has often been labelled the 'century of refugees' in European History. After the Balkan Wars and the Great War had already uprooted people across Europe, the Nazi rule in Germany and the Second World War led to unprecedented waves of forced migration and deportation. Violence during the first half of the century set millions of people in motion as refugees, victims of deportation or 'ethnic cleansing'. Those who survived war and genocide often kept moving as *displaced persons* for decades after 1945.[4] This marks the years between the 1930s and the 1960s, which we often predominantly see as the age of totalitarianism and the Shoah, as the years of the 'great migration', too, or the climax of a refugee crisis, which fundamentally reshaped Europe and left few countries in the world untouched.

The International Tracing Service (ITS) has been a key actor in managing this twentieth century migration crisis and has hence built the most important archive documenting the lives of refugees and displaced persons searching a way back to a regular life and peace.[5] Biographical accounts and documented itineraries of the civilian survivors of the Second World War and the Shoah, who moved across Europe and the world, possess the unique potential to reconstruct this refugee crisis – from displacement to resettlement – as a larger social process with empirical patterns and narratives simultaneously.[6] This advances our understanding of forced migration and its impact on groups, institutions, individuals, movement patterns, mobility hotspots, agency and decision-making as well as of its wider impact on societies in Europe and abroad. It also increases our understanding with regard to the different institutions and actors charged with the management of the situation or getting involved as stakeholders of specific groups of refugees.

While the ITS archive has so far mainly provided services for tracing individuals, which in turn have resulted in an indexing protocol geared perfectly to that purpose, this research use case aims to unlock its documents to a mass-data analysis leading to a visualization of refugee trajectories in order to model the great European refugee crisis based on thousands of individual lives and experiences. However, currently, the indexing of documents in the archive is performed on single document level rather than tying together various documents available on individuals. As a consequence, there is no way to retrieve all information on a specific person automatically or semi-automatically from the document database. This implies that the unique potential of the ITS archive to trace paths of refugees and/or displaced persons based on larger samples instead of beyond the individual or small-group level remains blocked.[7]

Given the size of the archive and its complexity, and also the groundbreaking research that would become viable on the grounds of matched person-relating data, explorative approaches to the challenge of transforming the indexing system are the first logical steps. This project covers two objectives: (1) sampling a representative cross-section of documents to enable an event-based reconstruction of patterns in which displaced

[4] Peter Gatrell, The Making of the Modern Refugee (Oxford 2013).

[5] Rebecca Boehling, Susanne Urban, Elizabeth Anthony, Suzanne Brown-Flemming, Freilegungen. *Spiegelungen Der NS-Verfolgung Und Ihrer Konsequenzen, Jahrbuch Des International Tracing Service 4* (Göttingen 2015).

[6] Henning Borggräfe, Hanne Leßau, Harald Schmid (eds.), *Fundstücke. Die Wahrnehmung der NS-Verbrechen und ihrer Opfer im Wandel,Fundstücke, Bd. 3 (*Göttingen 2015).

[7] Kathrin Flor and Verena Neusüs, Ein Archiv öffnet sich. Jahresbericht 2008 / The Archives open up : Annual Report 2008 (Bad Arolsen 2009).

persons migrated across time and space; (2) reconstructing individual itineraries, movement patterns etc. from a smaller sample of tracing and documentation files.[8]

The following research questions might be addressed: When unlocking the ITS archive as a collection of life-event mass-data with which we can build a time-space-model of great migrations, what general patterns and structures can we trace? Where did people go who left Europe before/in the wake of the Second World War? Can we improve our knowledge on why certain people were transported to particular places? Did people return to their town or city of origin after the war had ended or did they go somewhere else? Which connections between different groups on the move in relation to one another can we trace, including shared or diverting places and paths, actors and strategies?

4.2 Materials and Sources

The project will use a digitized sample of data from at least 10.000 persons from the collection of the International Tracing Service (ITS).

4.3 Procedure

The plan of action does not initially intend to modify ITS data structures. It aims at implementing a pilot project separately from the running database and serves as a testing ground. Necessary steps include (a) the identification of individuals in operable parts of the Central Name Index through semi-automated data processing to then link documents referring to an individual; (b) drawing a sample of 10.000 individuals in order to organize basic demographic information (name, surname, date of birth, place of birth, gender) in a database to gain an insight into *whom* is represented in ITS holdings; (c) gathering a sample of 1.000 TD files to extract similar demographic information as well as itinerary data; (d) geocoding and timecoding life-events that document movement and visualize the data in a Geographic Information System (GIS); (e) building an exemplary collection of narrative sources on selective subgroups to demonstrate the potential of linking qualitative documents (narratives) to life-event data.

This research case is part of the project *People on the Move* which is a collaboration between NIOD, Institute for War, Holocaust and Genocide Studies in Amsterdam, the Netherlands[9], the Institute of Migration Research and Intercultural Studies (IMIS) of Osnabrück University[10], Germany and the ITS[11].

4.4 Expected observations

Following large numbers of people on the move through Europe from the 1930s throughout the 1950s, and tracing changing patterns in their movements, means gaining an overview of spaces and moments and the way they 'stamped' and transformed the identities of the people concerned. To pursue this endeavour, the study of people on the move needs to go beyond those caught up in the Nazi persecution system and involves forced labourers, war refugees, and many other categories – including people in post-war DP camps which could

[8] Christoph Rass, 'Sampling Military Personnel Records: Data Quality and Theoretical Uses of Organizational Process-Generated Data', *Historical Social Research/Historische Sozialforschung*, Bd. 34 (2009) 172–196.
[9] Ismee Tames, see: http://www.niod.nl/en/ and http://niod.nl/en/medewerkers (last accessed 04 April 2016).
[10] Christoph Rass (http://www.chrass.de), Sebastian Bondzio (https://www.geschichte.uni-osnabrueck.de/bondzio_sebastian) (last accessed 04 April 2016); the IMIS team also includes Janis Panagiotidies, Sebastian Huhn and Olaf Berg.
[11] Henning Borggräfe.

be regarded as liminal spaces par excellence, for instance, for those waiting for a possible transfer across the ocean.

4.5 Results/Analysis

This research use case aims to contribute to our understanding of migration movements of European refugees before, during, and after the Second World War and the way general patterns and structures changed over time. The individual files of the ITS are ideal sources to execute this research as they allow to map more general migration movements and connect them to individual experiences.

5 Between Decision Making and Improvisation. Tracing and Explaining Patterns of Prisoners' Transfers through the Concentration Camp System

5.1 Purpose

This research case builds on recent studies on the role of IBM in the Holocaust, and in particular on the role of the so-called Hollerith-Abteilungen (Hollerith Department) in the Nazi concentration camps; assigned with keeping tabs on inmates through use of IBM's punch card technology. The main question we pose is how the SS in the years 1942-1945, via these departments, managed the transport of inmates through the concentration camp system that covered the whole of occupied Europe. When forced labourers were needed at a certain place, transports were organized from one camp to another. Can we discern patters in these transport and the related selections processes (from which – central or sub – camp)? Which type of prisoners were organised and which age groups and nationalities? Which professions were included, etc.). Furthermore, can we trace how these transport patterns changed over time? How did this system function during the last chaotic year of the Third Reich? Can these characteristics and changes be connected to the networks of leading SS officials, which according to the historian Karin Orth, consisted of 30 men, and their rotation in the concentration camp system? Or were decisions taken on a more central level in the SS-Wirtschafts- und Verwaltungshauptamt (SS-WVHA), which was founded in March 1942?[12]

5.2 Materials and Sources

A good starting point for this project is the WVHA-Häftlingskartei (a joint reconstruction project relating to the fragments, cf: https://portal.ehri-project.eu/units/de-002512-wvha_h%C3%A4ftlingskartei). Additional archives are likely to be found at 1) ITS Arolsen (all collections relating to Nazi concentration camp registries and transfer lists); 2) The archives of the Concentration Camp Memorial Sites in Germany, Austria, Poland, France, the Netherlands, etc.

5.3 Procedure

The basic procedure of this research case is to develop statistical models and correlations based on several of the above-mentioned features and attributes (age, occupation, nationality, etc.)

5.4 Expected observations

Digging into the pre-history of computing as related to the Holocaust, we expect to see patterns that point to newly uncovered selection procedures within the camp system. These will confirm and discover transport relationships between different camps in the camp system. Furthermore, because of the richness of the structured information involved, chronological changes in the camp system should be confirmed or become apparent. Finally, the work should enable new research on the making of the concentration camp system through management as well as balanced information on local or central decision-making.

[12] This project proposal is based on ideas of Dr. Edith Raim.

5.5 Results/Analysis

This research will demonstrate how digital methods can confirm existing understanding, which is mainly gleaned from anecdotal witness evidence and surviving SS-analysis, as well as develop new insights and relationships not covered so far but that are more obvious to statistical analysis rather than other types of archival research.

6 Archives and Machine Learning

6.1 Purpose

The main idea behind this research case is to investigate how digital methods might support archivists in the creation of interoperable and consistent descriptions of sources (metadata) and in the linking of sources. We intend to analyse the extent to which it is possible to automatically generate high-level collection description summaries in cases where only file-level descriptions are available, and to use predictive analytics to evaluate the consistency of existing collections.

6.2 Materials and Sources

The project will mainly rely on the existing data within the EHRI portal (https://portal.ehri-project.eu/) in order to understand and improve archival description work. In order to verify the description additional archival sources that are highly structured will be consulted such as the USHMM name database: https://www.ushmm.org/online/hsv/person_advance_search.php.

6.3 Procedure

In a first step, we intend to analyse recent approaches to predictive analytics and anomaly detection (such as Bayesian and Network reasoning) in order to understand how metadata is generated in the context of archives. A second step will be to develop procedures for automatically generating archival metadata from digitised file collections and creating semantically meaningful links between files and collections. We can rely on a range of recent approaches to automatically generate metadata and links from web pages using semantic extraction. Archives - with their deluge of collections - will, in the future, have to rely more on such computer-generated techniques in order to make effective use of their collections, augmenting but not replacing the efforts of archivists.

Our investigation will experiment with machine-learning techniques to generate collection descriptions and links between its constituent items. The second aim of the research would be to investigate existing descriptions and links in the EHRI graph store for consistency and interoperability. Expected techniques include information extraction and automated link generation as well as probabilistic graph models to describe link frequency and consistency.

We would include established procedures from machine learning to evaluate the meaningfulness of the automated metadata generation. For instance, files will be clustered using several metrics in order to understand how automatically generated centroids compare with manually assigned ones. The clusters will then be evaluated according to how dense and well-separated they are. Finally, a labelling process will be defined in order to understand how the clusters could be understood as new collections.

6.4 Expected observations

The first observations stem directly from the above described procedure. Content description as well as metadata development are not deterministic. Different individuals summarise collections in different ways, and the question will be whether a computer could simulate this. We expect to see discrepancies between human and machine, but also between human and human as well as machine and machine annotations that are common in such investigations. The novelty of the research will lie in publishing systematic summaries of these observations.

A second observation will be to what extent professionals working with and in archives and who are used to working with highly curated collections are willing to include probabilistic categorisation and cataloguing into their work processes. It will be interesting to see how this dynamic plays out.

In effect, this research will investigate the potential for algorithmically-assisted archiving of historical material. In order to manage the ever-expanding digital collections in many contemporary archives, a computer has to assume a role as a trusted intermediary and custodian of knowledge, with data retrieval and analysis algorithms determining what is knowable in an archive. Their role will increasingly be to decide what is part of an archive and how our understanding of it can evolve. By comparison, in traditional archives it is the archivists themselves who set the criteria for acquiring records according to their 'enduring value' (http://www.archives.gov.on.ca/en/about/archives_unboxed/archivist.aspx) to their organization. They curate their collections and help researchers discover files. How this model of trust can be adapted to encompass algorithmic processes is currently an unresolved question.

6.5 Results/Analysis

We expect to determine the principal feasibility of a computational approach to archival descriptions at scale and would evaluate this with A/B testing of a user group. The machine learning results will be evaluated using dedicated test data that includes known labels for clusters as well as metadata, as assigned by experts. The final aim would be move towards a gold-standard collection for archival metadata generation on the Holocaust. More importantly, however, we will evaluate whether the clusters support effective research reasoning with the collection. Will the automatically generated collection descriptions help the archival user understand the underlying files better and will they add to the research quality of the archival material?

In terms of the consistency of existing resources, we will generate comprehensive statistics using accuracy numbers generated by the machine learning algorithms. We expect to see distinct patterns that illuminate how humans and machines cluster collections.

6.6 Conclusion

The work aims to bring machine learning techniques to the organisation of archives. EHRI has by now accumulated a unique data set of structured information around archives that we can use to understand human and machine generation of metadata and potentially improve it. The research case investigates how archivists in the future can benefit from digital methods. For EHRI especially relevant is the use of graph and network reasoning that we can use to understand and improve the connectedness of our collections.

7 Networked Reading

7.1 Purpose

The purpose of this research case is to investigate how digital historiography can move from ad-hoc experimentation to systematic investigation. Its aims are two-fold: The first one is to understand the form historical sources need to take in order to be processed using digital methods. What kind of infrastructure do we need to extract meaning from them? The second challenge is how to apply these digital methods in such a way that the results are verifiable as well as reproducible. Here, we aim to find out how we can proceed with a networked reading of historical documents.

7.2 Materials and Sources

Part of the research in the project is how to transform existing unstructured information (mainly textual but also image sources) into research data sets, record the steps and define the infrastructure. In this sense, we work on derived data sets and on building new data sets. We are currently already investigating the US Holocaust Memorial Museums oral testimonies collections (https://www.ushmm.org/research/research-in-collections/overview/oral-history) and will expand our efforts to include more such as the NIOD's The Kingdom of the Netherlands (http://www.niod.nl/en/projects/enriched-kingdom) as well as Yad Vashem's archival materials.

7.3 Procedure

We will explore a number of data sets within the EHRI consortium - selected based on various distinctive features such as topic, format, time period, etc. - and systematically record the changes we need to apply to enable computational reasoning with them. If possible this should result in generic representations to which more than one digital method can be applied. Furthermore, we will explore how the preparation of the data for processing, such as the replication of missing entries through summary statistics or the dropping of entries with missing attributes, will influence the generation of historical knowledge. In a second step, we will define curation and preservation procedures for these collections that will make them accessible in the long-term and analyse how they can be meaningfully shared within the framework of larger initiatives such as DARIAH.

Based on the curated research collections, we will analyse which kinds of digital methods can be applied based on the profile of the underlying data; what is required to develop computational models of the data in a systematic rather than ad-hoc manner. This includes an in-depth investigation into the types of evaluation strategies that seem most applicable to the data and research at hand; as well as an analysis of how to gain insights from the data and what kind of questions can actually be answered by an investigation such as this.

For digital humanities, this research describes a grand challenge, since much of the current work in the field is characterised by the ad-hoc application of tools to gain quick insights on data sets, without embedding them in a larger framework of data creation and curation, computational modelling and evaluation. In the experiment, we seek to bring together the best practices of the tradition of historical interpretation with the best practices of the modern science of document and data analysis as well as social science practices of content analysis. From a digital humanities point of view, it is important that the field moves beyond being an assemblage of enticing and alluring methods, tools and procedures. This requires systematic investigations into how digital methods relate to digital data in history and Holocaust studies, and into the application of such methods in a way that ensures that

experiments can be repeated and thus understood and verified by the community as well as evaluated and peer-reviewed.

7.4 Expected observations

We hope to see as a first observation how Holocaust research material can be transformed into actionable research data sets. This will generally require a number of pre-processing steps as well as a focus on the kind of analysis we would like to do with them afterwards. There is a lot of preliminary work in related fields such as criminal history but little has been done in Holocaust studies. We will systematically describe preliminary steps to 'code' Holocaust documents using text mining and other techniques with important information about actors, networks, locations, etc. Hand-coding these documents is not only tedious but also leads to what is known in the social science as 'inter-coder reliability' issues. The need for hand-coding remains one of the biggest barriers for advanced computational analysis of historical and other cultural documents. We will investigate how far we can automate these processes without sacrificing the insights we can gain from the documents.

The first major contribution of this research will therefore be to systematically link human encoding techniques with computational ones to pre-label significant events, people, places, etc. What kind of techniques are already in use here to achieve an effective feature reduction - from automated clustering to summarising words into concepts using thesauri and other dictionaries. What kind of resources are available for Holocaust studies, and how can the scope and accessibility of such resources be broadened?

The second observation will be the reliability of existing digital methods when it comes to Holocaust research. Can they produce meaningful results so that we learn something about the Holocaust rather than about the (digital) condition of the data they use? Can we observe how methods 'think along' (Rogers) the material they work on? Finally, how can these methods be communicated to lead to results that are also accepted by the community? From these observations we expect to provide for the first time a systematic overview of the methodological resources available to digitally interested humanities researchers.

7.5 Results/Analysis

Our concluding objective is to develop a method of what we call 'networked reading and reasoning' of digital resources in Holocaust research. This idea of networked reading is distinct from the digital humanities concept of 'distant reading' by Moretti and others in that it focuses on identifying commonalities and links between documents while maintaining the close reading perspective of traditional historical enquiry. For such networked reading the amount of potentially relevant source material has to be considered; the semantically complex, 'messy' and dispersed nature of such sources; and the challenges arising from collaborative approaches to Holocaust research.

The second aim of networked reading is to discover larger trends that are expressed in the documents and how documents in a collection are networked by them. While we would like to avoid to use the term 'big history' here, we nevertheless believe that one of the major advantages of computational analysis is that it can produce more reliable longitudinal studies. This research use case wants to take a step back and think carefully about how we are to link and analyze historical documents in the digital age. How, for instance, can 'anomalies' that historians have discovered in the material (such as conflicting evidence against a commonly assumed theory) be put into the context of potentially larger investigations into trends and concepts?

The studies results will be threefold. Firstly, we will survey digital methods and develop an overview, secondly we will produce guidelines for coding historical collections, and finally we will present a procedure for transforming research material into research data. In a final step we will present how data can be linked to methods to enable network reading.

7.6 Conclusion

The research case adds to digital humanities a new systematic approach to applying digital methods to historical collections. In particular, we will develop systematic means of creating digital research collections from archival material that can be used by digital methods. Finally, we will embed the research in a larger attempt to create a new kind of research with digital methods that can be accepted by historians. In particular, we will look to develop new methods for networked reading of historical documents.

8 Conclusion

The research use cases presented in this deliverable aim – taken together – to facilitate the co-investigation and co-development of digital methods and tools between Holocaust researchers and digital specialists. Regarding the broad spectrum of historiographical topics, the diverse corpora of data and different methods that are addressed the prospects of success are promising. For every specific selected data collection the ethical and legal conditions that need to be maintained are different.

We presented six use cases in total, each of which would be a research project in its own rights. We concentrated especially on history research but also added the two other disciplines in EHRI. All historical use cases clearly concentrate on people and their relationships. The second, third and forth use case add to this an interest in movement of people; this probably reflects current interests with regards to migration. The final two use cases work on using advanced computational methodologies and technologies to enhance archival research that are on the one hand side useful to archival work on the Holocaust but are on the other hand also of more generic use. The next step for the WP will be to define which of the use cases the WP will concentrate on during the final three years of the project. We will take into consideration the availability of data as well as legal and ethical requirements. It should be noted, however, that we also foresee that new use cases will be added and the existing ones updated during these three years.

Currently, the WP is experimenting with initial computational methods for a digital historiography and will present these at the General Partner Meeting 2016. We have had some initial success with tests using entity extraction using archival finding aids from USHMM as well as survivor testimonials. Both tests showed promise as they will also allow us to compare the manual metadata work we are doing in EHRI with automated processes. Topic modelling as well as network analysis were used to understand common themes across these collections. Currently, the WP proceeds with its technical work in three working groups on (1) names and networks, (2) topic and language shifts and (3) geographies. Finally, the WP agreed to proceed with our digital methods work using virtual machines and iPython notebooks. This will allow for an easy integration of tools with data and avoids complications with platforms in order to establish open tools and methods.