Historical Research and Evidence in the Digital Age.
The Case for New Approaches in Holocaust Research

by Tobias Blanke, Conny Kristel

Abstract

Digital approaches to Holocaust research have led to a renewed interest in how researchers in the humanities work with material and use it as evidence for their work. Creating evidence by looking for connections and making links between events, people and places is key to all historical research. The most basic methods of generating evidence in history have changed with the digital transformation of archives and the scholarship linked to them. This paper investigates this digital transformation based on the experience of the European Holocaust Research Infrastructure (EHRI).

Introduction

Background: Historical Research

Identification, Provenance and Iteration. The EHRI Experience

Conclusion

_____

Tobias Blanke, Conny Kristel

Introduction

Digital approaches to Holocaust research have led to a renewed interest in how researchers in the humanities work with material and use it as evidence for their work.[1] While many have welcomed this, for others confusion reigns as to how digital humanities delivers new approaches that can be useful to other researchers in the humanities. This paper discusses this new material and evidence and considers how it can support historical research on the Holocaust. We developed three principles that guide the digital transformation of sources and evidence in Holocaust research. These three principles are the identification of sources, the development of trust and provenance and, finally, the question of the transformation of research processes and questions in an age of ever increasing numbers of sources. We derived these three principles by investigating current historical research outputs in postgraduate theses, which we present in the second section of this paper. The third section then presents these three principles as key to the digital transformation using the EHRI experience. Firstly, EHRI offers new means of critically identifying sources using graph databases. Secondly, EHRI has explored how the principles of source criticism and provenance, which are central to historical and archival theory and methodology in Holocaust research, can be maintained in the digital age. Thirdly, we discuss the logic of ever larger datasets. In our conclusion, we propose concrete steps which can teach and enable historians to work with digital evidence. The digital transformation makes us question again how we deal with evidence in general, which is of the utmost importance in Holocaust research.

In the next section, we will look at the actual practice of qualitative historical research and its interaction with evidence. This will help us develop the three core principles that guide our investigation of the EHRI experience of digital transformations. We will focus on one of the core activities of historians, i.e. searching for information and transforming it into evidence. Especially in a field such as Holocaust studies, working with and justifying evidence continues to challenge research.

---

[1] *Digital evidence: selected papers from DRH2000, Digital Resources for the Humanities Conference. University of Sheffield, September 2000*, eds. Marilyn Deegan, Michael Fraser and Nigel Williamson, (London: Office for Humanities Communication, 2001).

Background: Historical Research

Discussing digital evidence in Holocaust research has to start from a general discussion of the state of the work with sources in Holocaust research. In particular, we need to consider the general state of historical methodology and how it is studied and taught. To analyze the practices of historical methodology, we decided to look into recent postgraduate theses. Compared to other monographs in history, they have to follow stricter templates to be accepted for examination. Among the requirements is a discussion of methodology and development of an explicit research question. Using these postgraduate theses as a case study, we investigated the state of evidence and the uptake of different methods and approaches (such as oral history, close reading, discourse analysis etc.) in Holocaust research. Studying theses should give us insight into research practices of master students and junior scholars, and indirectly also in the way they are, or have been, trained at universities. To this end, we have analysed twelve recent theses, six Masters theses and six PhD dissertations in Holocaust research, which shed a light on the status quo of the practice of historical work.

We have looked at the section, usually the introduction, in which the authors are expected to discuss their research methodology and/or research procedure.[2] Most authors began the section with an exposé about their research question in relation to the *status quaestionis*. The research question is presented as guiding the strategy, the sources and the method. The famous French historian Lucien Febvre (1878-1956) phrased the historical working with sources as follows in 1933: "A historian does not simply wander around the past as a buyer looking for old rust but he assumes a precise plan in his mind, a problem to be solved, a working hypothesis to be verified."[3] The authors in the sample stated that their research was guided by the research question, but that this question was also transformed in the process of conducting research. Historical research is an

---

[2] Authors A, B, C, H, I, J have written a PhD dissertation; authors D, E, F, G, K, L a Master thesis.

[3] Original quote: "l'Historien, qui ne va pas rôdant au hasard à travers le passé, comme un chiffonnier en quête de trouvailles, mais part avec, en tête, un dessein précis, un problème à résoudre, une hypothèse de travail à vérifier." (Lucien Febvre, "De 1892 à 1933. Examen de conscience d'une histoire et d'un historien. Lecon d'ouverture au Collège de France, 13 décembre 1933," *Combats pour l'histoire*, ed. Lucien Febvre, (Paris: Librairie Armand Colin, 1992, digital ed.,), 3-18; 8. Accessible via:
http://classiques.uqac.ca/classiques/febvre_lucien/Combats_pour_lhistoire/febvre_combats_pour_histoire.pdf

iterative process. In our case study, Author H explicitly mentioned that she had revised her research question during the process: "I intended to write a concise history of x's wartime operations, but the primary sources I had uncovered told a more complex story." Seeking evidence in history is first and foremost an iterative process through an ever-increasing amount of sources that enable new research questions and lead to ever larger sources to (re-)consider.

The second major component of evidence-based research is the identification of sources. In our case study, the authors noted which sources they had used and, especially in the case of the PhD authors, reflected on the *scope* of the sources.[4] However, it is often more difficult to understand the exact functioning of particular sources. In ten theses we were not able to find an explanation for the use of particular sources that went beyond their sheer identification. Only two PhD authors explain their research strategy with regard to the sources. Author C, who wrote about a family history, made his research strategy explicit. For every generation, he carefully examined a series of the same themes. Author J analyzed an intellectual's thinking about Western culture through the lens of three oppositions. After reading the introduction, the reader knows which question(s) the researcher aims to answer, how the thesis connects to existing research and which source material the researcher has selected and consulted. For sources to be turned into evidence, however, the researcher is required to describe the provenance of these sources, beyond their pure identification.

None of the authors explained in their publications how they found their sources. Author B at least mentioned a "snowball method," but offered no further details, such as which document or resources were the beginning of the snowball; nor did she reflect on the potential problems with this method. In their *History Manifesto* (2014), the American historians Jo Guldi and David Armitage argue that there is a bright future for historians, if they not only investigate "forgotten stories," but also take on the role of "arbiters of data for the public." According to Guldi and Armitage, historians are excellently positioned as teachers of a critical approach which will only become increasingly more necessary in the digital age with an information overload.[5] If historians want to become the future teachers of the critical approach, this lack of transparency in

---

[4] Author A, 18; author B, 30, 31, 33; author C, 29, 30; author I, 14; author J , 23; author L, 5.
[5] Jo Guldi and David Armitage, *The History Manifesto*, (Cambridge: Cambridge University Press, 2014), 88-116; 113.

terms of the methodological underpinnings of their work is something which needs to be addressed.

Our brief investigation and short case study demonstrate that forming a historical argument, or rather the way in which a historian working with empirical material transforms sources into evidence, is complex and intricate. Even the term "sources," which historians commonly use to refer to their documents, can be a misleading metaphor. It implies "the possibility of an account of the past which is uncontaminated by intermediaries." As historian Peter Burke has suggested, it might be preferable to speak of "traces" instead.[6] The process of selection and interpretation of these traces is based on rules which are inherent to the historical method. The latter is a specification of the generic scientific method. Historians ask different questions and cannot perform tests. This being said, many of the rules that historians are expected to apply in order to find answers that are logical and supported by evidence, are generic for scientific reasoning. Specific for historians is the scholarly principle of source criticism, which aims to critically assess various aspects of a source. Even in generously referenced publications this process is usually not very transparent, nor can it easily be reconstructed, which from a scholarly point of view is problematic. Ideally, historians should allow their colleagues access to the "kitchen" so that they can understand – or even witness – how the publication has been prepared

Analyzing historical statements and claims and tracing them back to the sources on which the author claims they rest is key to any historical research, but especially so for contested and often politically sensitive areas such as the Holocaust. Here, researchers tend to rely on not just on their own but also their colleagues' work, given that the Holocaust provides a vast amount of traces and sources. If historians do not work together, as the British historian and specialist of the history of the Third Reich Richard Evans has argued, "it would be completely impossible for new historical discoveries and insights to be generated."[7]

---

[6] Peter Burke, *Eyewitnessing: The Uses of Images as Historical Evidence*, (London: Reaktion Books, 2001), 13. Burke is leaning on a suggestion of the Dutch historian Gustaaf Renier (1892-1962), who opted for the term "traces" in his book *History, its Purpose and Method*, (London 1950).

[7] Richard J. Evans, *Lying about Hitler: History, Holocaust and the David Irving Trial*, (New York: Basic Books, 2001), 18.

For Holocaust research, it is even more important that researchers work together, as they are continuously challenged by those who seek to deny, or at least diminish, the systematic mass murder that occurred. Sources, traces, and their corresponding evidence on the Holocaust are routinely challenged by those whose agenda it is to reduce its historical importance and uniqueness. One of the most famous incidences relating to this specific challenge occurred in the late 1990s, when David Irving filed a libel suit against the American historian Deborah Lipstadt and Penguin, the publisher of the British edition of her book on Holocaust-denial*, Denying the Holocaust* (1993). In this book, Lipstadt calls Irving "one of the most dangerous spokespersons for Holocaust denial." The lawyers of the defendant commissioned four professional historians to write expert reports on specific elements of the defense. The defense in itself was challenging, for in accordance with the English law of defamation, it was up to the defendant to prove that the defamatory statements (by Lipstadt) were true.[8]

Evans was one of the four experts. He was asked to go through a sample of Irving's work, which until then was considered by some prominent critics as a relatively good work of amateur history, and deliver a report "on whether or not Lipstadt's allegation that he falsified the historical record was justified."[9] Evans subjected Irving's work to a detailed dissection of historical statements and claims, using a series of investigations that exposed how evidence and traces should be treated in Holocaust research.

Irving lost the trial. In *Lying about Hitler*, which is based on Evans' expert report and on his experiences during the trial, Evans demonstrated that Irving's argument failed to withstand the test of source criticism and evidence-based research. *Lying about Hitler* is not so much a theoretical essay, but more a detailed analysis of the methods Irving used in his attempt to turn information into evidence. Evans demonstrated clearly how Irving worked with his sources and where he was blatantly not only ignoring, but also bending, the rules. The judge ruled that Irving had "misrepresented and distorted the evidence which was available to him" to fit his political agenda. The judgment was, as Evans rightly

---

[8] *Ibid.*, 28-9. Apart from Evans, the defense commissioned Christopher Browning, Peter Longerich and Robert-Jan van Pelt to prepare expert reports. These can be found via: https://www.hdot.org/trial-materials/witness-statements-and-documents/.
[9] Evans, *Lying about Hitler*, 30.

writes, "a victory for history, for historical truth and historical scholarship."[10] It also demonstrated, however, how painstaking and time-consuming it is to take apart an argument once it has been neatly put together in a publication.

With this particular use case in mind, we will take a closer look at the historical method in practice. The British historian and philosopher Robin George Collingwood (1889-1943) qualified history as "a certain kind of organized or inferential knowledge."[11] Taking this statement as a point of departure, and recognizing that historians one way or the other work with evidence, the rest of this paper reflects on the organizational dimension of the research practices of historians in light of the digital transformation of historical research, with a special focus on research into the Holocaust. In particular, we investigate the key components of iteration through ever-increasing sources, identification and provenance as those elements that define historical work with traces and sources. Our presentation is based on our experience of the digital transformation of historical research in the framework of the European Holocaust Research Infrastructure project (EHRI).[12] Before we go into the details of iteration, identification and provenance we provide a brief overview of the state of the digital transformation of historical research.

Identification, Provenance and Iteration. The EHRI Experience

Nowadays, historians who want to work with sources and traces in archives generally do not have a choice whether to engage with the digital. From our experience in EHRI, we know very well that only a small part of archival collections in heritage institutions is available in a digital format, but searching for particular files or information is generally done via digital means. We do not want to say that all archives will be fully digital anytime soon, but there are already enough archives which support digital engagement, because it makes sense, not just in order to deal with ever increasing volumes of records, but also because access to even the smallest record is much easier and more flexible by digital means. EHRI is attempting to integrate what has been digitized about the

---

[10] *Ibid.*, 265.
[11] Robin George Collingwood, *The Idea of History*, (Oxford: Oxford University Press, 1961), 252.
[12] http://www.ehri-project.eu

Holocaust and provide access to the rest. How much significant material on the Holocaust will ever be digitally available remains to be seen, but more and more the perception of Holocaust material will be based on what is digitally available. This is a common experience of digital transformation for all archives and research in them with which we have worked.[13]

Identification of sources

Archives are digital organizations and businesses, as the whole world has become digital and the humanities and history are part of it. Realizing this is the real meaning of digital transformations. It is the realization of a digitized world that drives a new understanding of some of the core concepts humanities research engages with. To us, one of these concepts is the idea of evidence-based research on sources. Traditional humanities scholarship does not have to be reduced by the digital transformation, but rather "could raise the critical standard for how we read all kinds of evidence."[14] We argue that even the most basic methods of generating evidence in history by identifying sources have changed, because we are confronted not with human-indexed collections in archives, but with machine-indexed digital ones. Just like in the pre-digital age, critical evidence from collections required an understanding of archival processes and the metadata work of archivists; in the digital age historians need to understand how computers identify sources.

Archival documents are presented to researchers not at the moment of digitization, but at the moment we use any kind of search engine to identify and access them. The digitally transformed understanding of working with sources, therefore, has to start with the most basic ways of integrating them through searching digital archives - taking for granted that these digital archives might not be based on sources from a single repository but may integrate many. How many repositories are integrated only matters to the digital search in so far as it raises the requirements for the infrastructure. The most basic step of the identification of sources is today, therefore, already a complex digital task that

---

[13] Sheila Anderson and Tobias Blanke, "Taking the Long View: From e-Science Humanities to Humanities Digital Ecosystems" in *Historical Social Research/Historische Sozialforschung* 37/3 (2012): 147-64.
[14] Tim Hitchcock, "Confronting the Digital: Or How Academic History Writing Lost the Plot" in *The Journal of the Social History Society* 10/1 (2013): 9-23; 20.

requires careful consideration of what kind of knowledge is won, but also lost, by approaching sources this way. Furthermore, it could also imply a new kind of publication of research that provides insight into how the search and find process has been undertaken.

In the excitement for more advanced digital history methods, sometimes the importance of simply searching through databases and the new ways of evidence this provides is forgotten. Not so with Bruno Latour et al, for whom the abilities for researchers to work through digital databases offer completely new opportunities of researching the social order: "[W]e wish to consider how digital traces left by actors inside newly available databases might modify the very position of those classical questions of social order."[15] For the sociologist, databases are exciting as they provide completely new networks of actors' traces. The research can move from the individual to her group and back without effort. The situation is not different in archival research. Databases allow researchers to effortlessly move between document and collection levels, and identify both.

Taking the importance of the critical identification of sources seriously motivated EHRI to commit to a graph database, which complements standard access and identification facilities. We added a complex system of navigating the generics and specifics of evidence in collections. We use traditional faceting, but also allow for more complex queries to a specific part of the target records. This leads to a close integration between searching and hierarchical browsing of country, institution and collection metadata in the EHRI portal. At any point in time, a user can choose how to proceed with the identification process. We think this is a good compromise between full search and an expansive and intensive view of the sources.[16]

The relation between full search and an expansive and intensive view is extensively explored in the axiomatic theory of information retrieval.[17] According to this theory, searching through archival collections is progressing by excluding

---

[15] Bruno Latour, Pablo Jensen, Tommaso Venturini, Sébastian Grauwin and Dominique Boullier, "'The whole is always smaller than its parts' – a digital test of Gabriel Tardes' monads" in *The British Journal of Sociology* 63/4 (2012): 590-615; 591.

[16] Tobias Blanke, Michael Bryant and Reto Speck, "Developing the collection graph" in *Library Hi Tech* 33/4 (2015): 610-23.

[17] Th.W.Ch. Huibers, *An Axiomatic Theory for Information Retrieval*, (Utrecht: PhD Dissertation, 1996), 58.

evidence that is not about what the researcher is looking for. A typical researcher is rather looking for any evidence before concluding the search. Judges, and other agents of the law, consider digital evidence differently. One needs to be able to rely on it in court: "Digital evidence is information stored or transmitted in binary form that may be relied on in court."[18] The systematic identification of sources and analysis of their evidence is called (digital) forensics.

Dan Edelstein demonstrates an example of digital forensics in literary history. By using the research API of JSTOR, he is able to discover trends in enlightenment scholarship and literature review. Key to his work is the ability to download quality data sources that are trusted on a particular topic.

> Because [research] is ultimately about assessing the quality of other people's arguments, it will and should remain a fundamentally qualitative exercise. The question is, are we always assessing the right works? Are we missing important trends? […] What data mining can offer, I suggest, is a broad yet detailed backdrop that helps guide our analyses of secondary sources.[19]

Only with practiced critical digital forensics can the principle of source criticism be maintained in the digital age. In archival research situations and in critical (digital) historiography in general, establishing the trustworthiness of sources is crucial. As the case of Evans versus Irving has demonstrated, it requires elements of both archival and historical theory and methodology, and has for a long time been the main source of collaboration between historians and archivists. Historians bring the traditional scholarly principle of source criticism to the table, whereas archivists focus on the archival principle of provenance. Both principles draw attention to the importance of providing contextual information for historical records.[20]

---

[18] "Digital Evidence and Forensics," in *National Institute of Justice website*, April 14, 2016. Accessible via: https://www.nij.gov/topics/forensics/evidence/digital/Pages/welcome.aspx (accessed October 17, 2017).

[19] Dan Edelstein, "Enlightenment Scholarship by the Numbers: dfr.jstor.org, Dirty Quantification, and the Future of the Lit Review" in *Republic of Letters. A Journal for the Study of Knowledge, Politics, and the Arts* 4/1 (2014): 1-26; 2.

[20] Katharina Hering, "Provenance Meets Source Criticism" in *Journal of Digital Humanities* 3/2 (2014): see http://journalofdigitalhumanities.org/3-2/provenance-meets-source-criticism/

Provenance and trust

Source criticism is a rather general critical approach of "sources" which is not exclusively applied by historians, although they are emphatically taught this principle as central to their discipline. The historian is trained to ask his or herself when, where, by whom, and for whom the source was made, and to critically evaluate the content, taking into account the context of the source in the broadest sense. Historians hereby mainly benefit from the archival concept of provenance, as it provides contextual information. The principle of provenance refers traditionally "to the individual, family, or organization that created or received the items in a collection."[21]

Historians seem concerned about the future of the related principles of source context and provenance in the digital age. Should and can they be maintained in a revised sense – and if yes, what are the contours of such a revision - or should we discard them altogether and maybe introduce new principles that suit digital sources and collections? The American information scientist, Katharina Hering, seems unwilling to throw the principles of provenance and context overboard. She argues that "the tradition of source criticism combined with a broadened understanding of provenance can support archivists, historians, librarians, digital humanists, and others with developing a set of questions and a vocabulary that can aid the analysis and description of digital collections."[22]

The archival principle of provenance is not just seen by the above investigated postgraduate researchers to be key to transforming sources into evidence. It produces trust. According to information scientists Wendy M. Duff and Catherine A. Johnson, in archival research situations sources can be trusted because of the "provenance method."[23] "Provenancial properties" are one aspect that defines the process that transforms archival "information into evidence"[24] for researchers. However, it has proven to be surprisingly difficult to maintain

---

[21] *Ibid.* The principle of provenance also suggests that records originating from the same source should be kept together, and should not be interfiled with records from other sources, to preserve their context.

[22] *Ibid.*

[23] W.M. Duff and C.A. Johnson, "Accidentally found on purpose: Information-seeking behavior of historians in archives," in *Library Quarterly* 72/4 (2002): 472–96.

[24] Joshua Sternfeld, "Archival Theory and Digital Historiography: Selection, Search and Metadata as Archival Processes for Assessing Historical Contextualization" in *The American Archivist* 74/2 (2011): 544-75; 551.

provenance digitally. Provenance is generally lacking online, where most of the digital evidence is found. The absence of "provenancial properties" in digital evidence is problematic for research. Archival practices do not always meet the principles taught to students of archival science. Mark Vajcner has called on archivists and archives to be "more active in ensuring that contextual information is linked to digitized materials."[25]

To keep provenance information in digital works, we have decided in EHRI to take each digital information object as a different one from its canonical item depending on the context it appears in.[26] This is our digital transformation of the "respect des fonds" which emphasizes the importance of maintaining the original structure of a collection. In this sense, the canonical item 'Israel|Yad Vashem|Jan Karski' is different from 'A's Research|Yad Vashem|Jan Karski' or 'C's Research|Monograph B|Yad Vashem|Jan Karski.' In EHRI, the identity of an item in a virtual collection is thus determined by the route to it that forms its context. This is our way of transforming provenance for digital archives and ensuring that archival information can become evidence for researchers.

Provenance remains the determining factor of how evidence is combined from sources in digital archives of the Holocaust. Provenance, however, also speaks to the larger question of trust, which is fundamentally digitally transformed by the unprecedented power of algorithms over our research processes. How can we trust the sources that are presented to us by undecipherable algorithms, which read and decode them for us? There remains an uneasiness in research with the way modern search engines et al. let the archival objects speak to us. At the same time, we have no choice. Historical sources grow quickly and already consist of millions and, likely, billions of documents.[27] What can be done with all these documents? In a perfect world perhaps experts would read the collection and index it for perfect retrieval by others. This is not possible with hundreds of millions of documents. The alternative has become to let very fast computer clusters read them in their particular way and extract all the relevant documents.

---

[25] *Ibid.*
[26] Blanke, Bryant and Speck, "Developing the collection graph."
[27] Tobias Blanke and Andrew Prescott, "Dealing with big data" in *Research Methods for Reading Digital Data in the Digital Humanities*, eds. Gabriele Griffin and Matt Hayler, (Edinburgh: Edinburgh University Press, 2016), 184-205.

But there is a growing unease with this alternative of fast but also non-transparent computing power operating in a black box. Frank Pasquale even calls this world the black box society.[28] The unease is also reflected in a very pessimistic and provocative intervention from three Dutch historians, who surveyed almost 300 colleagues in the Netherlands and Belgium about their online search behaviour. The results showed that scholars mainly search for text and images and that general search systems (as Google and JSTOR) are predominant. Most of the surveyed scholars searched with keywords, and they hardly ever used advanced search options to iterate through sources. The authors argued that Google introduced a black box into digital scholarly practices, and voiced concern that scholars will become "increasingly dependent" on such black boxed algorithms. Therefore, they recommend "a reconsideration" of the academic principles of provenance and context.[29]

It is within this context that media historian Andreas Fickers called for a "digital historicism" (instead of a "digital escapism"). According to Fickers, we need new historical practices to confront the danger of black boxes for research:

> [F]uture historians cannot escape the productive confrontation with the new technical, economic and social realities of the digital culture. Instead of digital escapism and methodological conventionalism the discipline of history is rather in need of a new digital historicism. This digital historicism should be characterized by collaboration between archivists, computer scientists, historians and the public, with the aim of developing tools for a new digital source criticism.[30]

Fickers perceives a parallel with the 19th century, which saw the emergence of history as an academic discipline. This led to huge editorial projects which were, more often than not, inspired by political and ideological interests (such as

---

[28] Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information*, (Cambridge: Harvard University Press, 2015).

[29] Max Kemman, Martijn Kleppe and Stef Scagliola, "Just Google It" in *Proceedings of the Digital Humanities Congress 2012. Studies in the Digital Humanities*, eds. Clare Mills, Michael Pidd and Esther Ward (Sheffield: HRI Online Publications, 2012). http://www.hrionline.ac.uk/openbook/chapter/dhc2012-kemman.

[30] Andreas Fickers, "Towards A New Digital Historicism? Doing History in the Age of Abundance" in *Journal of European Television, History and Culture* 1/1 (2012): 19-26.

nation-building).[31] All historians are witnessing today this paradigm shift from a "culture of scarcity to a culture of abundance." This has resulted in a "crisis of historical practice."[32] The next section is dedicated to this culture of abundance and sources at scale.

Sources at scale

Our case study above has identified as a third principle of historical approaches to evidence the iteration of research questions based on new sources, which leads to an ever expanding cycle of questions and sources. Recognizing this need, EHRI is, at its core, an attempt to bring historical data together to create an ever bigger linked-up dataset. It is an integrating infrastructure for researchers to work through other infrastructures containing sources. As such, EHRI is not simply a new instrument for historians, but also changes their work fundamentally. They are pushed into methods that reflect these ever larger and integrated datasets. Critical infrastructure studies have been aware of these effects for a long time.[33] Infrastructures enforce a particular view on the evidence. Traditional relational databases, for instance, sort what they represent into tables – not unlike Excel spread sheets. They also require their data to be combined using joins, which rely on unique identifiers given to each record in the data set. Databases imply a logic of joining and linking to create ever larger datasets.

Once evidence is in a database, there is in principle no limit regarding links to other databases and the integration of new datasets. This expands the sources and the questions that can be asked. Because integrating information is so important for the success of an organization's digital infrastructure, the archive will be involved in a permanent effort of standardization of data and metadata that is in use in this organization. Beyond individual archives, integrating infrastructures such as EHRI bring together evidence from multiple sources using computational reasoning that operates efficiently with sets of evidence. These combined sets then produce even bigger sets, and so on, ad infinitum. We

---

[31] *Ibid.*, 26.

[32] Roy Rozenzweig, "Scarcity or Abundance? Preserving the Past in a Digital Era" in *American Historical Review* 108/3 (2003): 735-62; 737.

[33] Rob Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures and their Consequences*, (London: Sage, 2014).

can merge ever more evidence with different sources: databases, networks of actors, as Latour et al envisioned.[34]

This expansion of evidence by its integration and merging can have two main issues. The most straightforward one is that we get lost in information overload. Because we seem to have everything, nothing matters anymore. The Linked Data vision suffers from this according to our EHRI experience.[35] A few years ago it was the next big thing, but as many semantic web innovations it has never really taken off, apart from niche areas in digital libraries, etc. Linked Data promised a universal standard to connect every digital piece of knowledge, not just in a single organization, but across the web, to every other piece of knowledge. Everything can be linked in a radical open world assumption. An evidence ecosystem develops from a "language of linking."[36] It quickly became clear, since this original vision, that forever expanding linkability limits us too. Neither humans nor machines have shown much appetite for infinite linkability. Smaller, more compact, visions of data integration have survived. EHRI has, therefore, committed itself to graph databases. We have discussed the challenges of 'linkability' in EHRI extensively elsewhere.[37] Here, we are concentrating on the second issue of the logic of ever larger datasets. The research questions that they enable lead to computing methods that can deal with the scale of the data assembled. The problem seems to be that large-scale data-driven methods tend to eliminate the development of specialized research questions in the first place, as they move from traditional statistics to big data techniques. This means that the cycle of iteration of sources to develop new research questions to find new sources, etc. is interrupted and replaced by a collection of sources on everything on which we can get our hands.

Donald E. Knuth is maybe the most famous godfather of computer science. For him,

---

[34] Latour, Jensen, Venturini, Grauwin and Boullier, "The whole is always smaller."

[35] Tobias Blanke, *Digital Asset Ecosystems: Rethinking Crowds and Cloud* (Oxford: Elsevier, 2014).

[36] Georgi Kobilarov, Tom Scott, Yves Raimond, Silver Oliver, Chris Sizemore, Michael Smethurst, Christian Bizer and Robert Lee, "Media Meets Semantic Web: How the BBC Uses DBpedia and Linked Data to Make Connections," in *The Semantic Web: Research and Applications. ESWC 2009. Lecture Notes in Computer Science*, eds. Lora Aroyo et al., (Berlin, Heidelberg: Springer, 2009), 723-27.

[37] For an extensive discussion of the challenges of "linkability" in EHRI, see: Blanke, Bryant, Reto and Speck, "Developing the collection graph."
http://www.emeraldinsight.com/doi/abs/10.1108/LHT-07-2015-0070

[s]cience is knowledge which we understand so well that we can teach it to a computer; and if we don't fully understand something, it is an art to deal with it. […] [T]he process of going from an art to a science means that we learn how to automate something."[38] Computing science is, therefore, about the tension to automate processes using digital means and not being able to do so, because we fail to fully understand the processes. In this sense, a computational approach to collecting and processing evidence would be a science if we could learn to automate it. Until then, it remains an art. Computational history is such an art.

Traditionally, the art of processing of digital evidence in history computationally took its inspiration from the social sciences, especially fields like sociology or psychology. For computers, processing digital evidence equates to solving an equation of the form $y = f(x)$. If x is the digital trace or source, then solving f leads to a decision y. So, if x would be a new revolutionary trace that Cesar never left Rome, then f could lead us to the conclusion y that he also never crossed the Rubicon. In statistics, there are a limited number of f to choose from, and additional parameters (so-called coefficients) are learned to fit this f to conclude from x that y. In statistics, this process is called forming a parametric model, which is

a learning model that summarizes data with a set of parameters of fixed size (independent of the number of training examples).[39] The model does not change, no matter how much data one throws at it, because the function f is limited from the beginning. This compares to nonparametric models, which are 'good when you have a lot of data and no prior knowledge, and when you don't want to worry too much about choosing just the right features.[40]

Nonparametric models seek to learn f and are able to generate it by themselves.

---

[38] Donald E. Knuth, "Computer Programming as an art" in *Communications of the ACM* 17/12 (1974): 667-73; 668.
[39] Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed., (Upper Saddle River: Pearson Education, 2010), 737.
[40] *Ibid.*, 757.

Parametric models of traditional statistics as used in the social sciences have been the first to inspire the analysis of digital sources in history. Martin Frické describes them as curve-fitting, or finding f. Here, the "task for science is to draw a curve (…) which will successfully anticipate the location of future or unknown data points'.[41] Such 'confirmatory data analysis," as is typical in social sciences and statistics, works with a prior hypothesis to examine and evaluate sources and create evidence. It has established theories and methods and the examples in history which take up this approach concentrate often on the domains where exact numerical calculations are possible. Processing digital evidence in history remains a social science here. History has known many examples of such traditional statistics approaches – often in economic history.[42]

Big data has generated enthusiasm for nonparametric models, and more and more data was inspired by a famous 2001 paper by Michele Banko and Eric Brill. It set the agenda for the big data hype and rationale. They demonstrated that calculating through digital evidences digitally gets more accurate by throwing more data at it.[43] Since then, the idea that "It's not who has the best algorithm that wins. It's who has the most data"[44] has been repeated multiple times as the magical unreasonable effectiveness of data. It has created the big data hype also in research and history.

For historical research, we are often limited to a particular kind of effectiveness of data, which are unsupervised methods. History in general, and Holocaust research in particular, lack larger annotated training collections that would allow it to apply supervised machine learning algorithms, which is why most existing large-scale textual methods in history use unsupervised techniques such as the clustering of documents or topic modelling. The problem with unsupervised methods is that they can almost be too easy to draw conclusions on. Representing documents as collections of words remains a brutal oversimplification of human

---

[41] Martin Frické, "Big Data and Its Epistemology" in *Journal of the Association for Information Science* 66/4 (2015): 651-61; 653.
[42] Donald N. McCloskey, "Does the Past Have Useful Economics?" in *Journal of Economic Literature* 14/2 (1976): 434-61; D.C. Coleman, "History, Economic History and the Numbers Game" in *The Historical Journal* 38/3 (1995): 635-46.
[43] Michele Banko and Eric Brill, "Scaling to very large corpora for natural language disambiguation" in *ACL '01 Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, (2001), 26-33.
[44] This quote has been attributed to Andrew Ng, professor in Robotics at Stanford.

language. For political history, the professor of economics at Stanford University Matthew Gentzkow and his colleagues in Finance and Econometrics at the University of Chicago Booth, summarize "[t]hat all automated methods are based on incorrect models […] also implies that the models should be evaluated based on their ability to perform some useful social scientific task."[45]

Unfortunately, the inclusion of validation into an analysis is not often the case in digital history. As an example, consider the popular method of topic modelling. Not one of its digital history examples we found actually validated the results. To this end, the researchers would have needed to first define "gold topics" in the domain they would like to model. Then, the topic modelling results would have had to be evaluated against these "gold topics" to understand how stable the topic model is. This is particularly important to topic models, as they heavily depend on a series (random) initialization steps. A model would be unstable if these would have a significant impact on the model's behavior. Also, topic models, like most machine learning operations, depend on so-called hyperparameters, which are set heuristically before the modelling can start. For topic modelling this is, for instance, the number of topics we would like to investigate. All this would require a number of validation steps, which are completely missing from most examples in digital history we have seen.

An unsupervised model like topic modelling can be tempting, as it suggests results without additional input and human effort. However, these topic models can be too suggestive and require careful intervention by humans. Even "unsupervised" techniques are effectively a human-computer assemblage.[46] For topic modelling to deliver digital evidence this requires a human to define gold topics for at least a part of the documents to be summarized into topics. In EHRI, we completed this exercise with an oral history collection of Holocaust testimonials. We chose 1,880 documents and split them up into further sub-documents or paragraphs of 500 words, as they were fairly unstructured interviews. We ran 20 topics against the whole collection and got topics such as:

---

[45] Matthew Gentzkow, Bryan T. Kelly and Matt Taddy, "Text as Data," February 15, 2017, available at http://dx.doi.org/10.2139/ssrn.2934001, 4.
[46] Claudia Aradau and Tobias Blanke, "Data-security assemblage: Knowledge and critique" in *Big Data & Society* 2/2 (2015): 1-12.

- *Topic 3*: school jewish did went time years remember children didnt yes high really know friends religious little parents kids home father

- *Topic 6*: germans going german train went just came time got saw soldiers didnt like army day russians started days took war

- *Topic 8*: camp people did know auschwitz didnt just work prisoners like time day yes ss got came camps barracks saw concentration

- *Topic 9*: jews people germans war russian poland polish army jewish russians came time did russia knew soviet hungarian started german hungary

- *Topic 10*: mother father went came family brother sister years got war died children did parents married didnt time lived husband left

- *Topic 11*: si nie na ja je tak po byo da bo jak tego eh jest mi ale te bya tym se

- *Topic 16*: people ghetto germans came didnt know went took jews place did going time like away jewish used killed work come

- *Topic 17*: food like little got bread used know water didnt eat day just gave took dont people piece work came big

While useful, the topics clearly require further careful interpretation by Holocaust experts, which demonstrates the additional work unsupervised methods such as topic modelling require. Figure 1 visualizes the 20 topics using the t-distributed stochastic neighbor embedding dimensionality reduction (REF).[47]

---

[47] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-SNE" in *Journal of Machine Learning Research,* 9/11 (2008): 2579-2605.
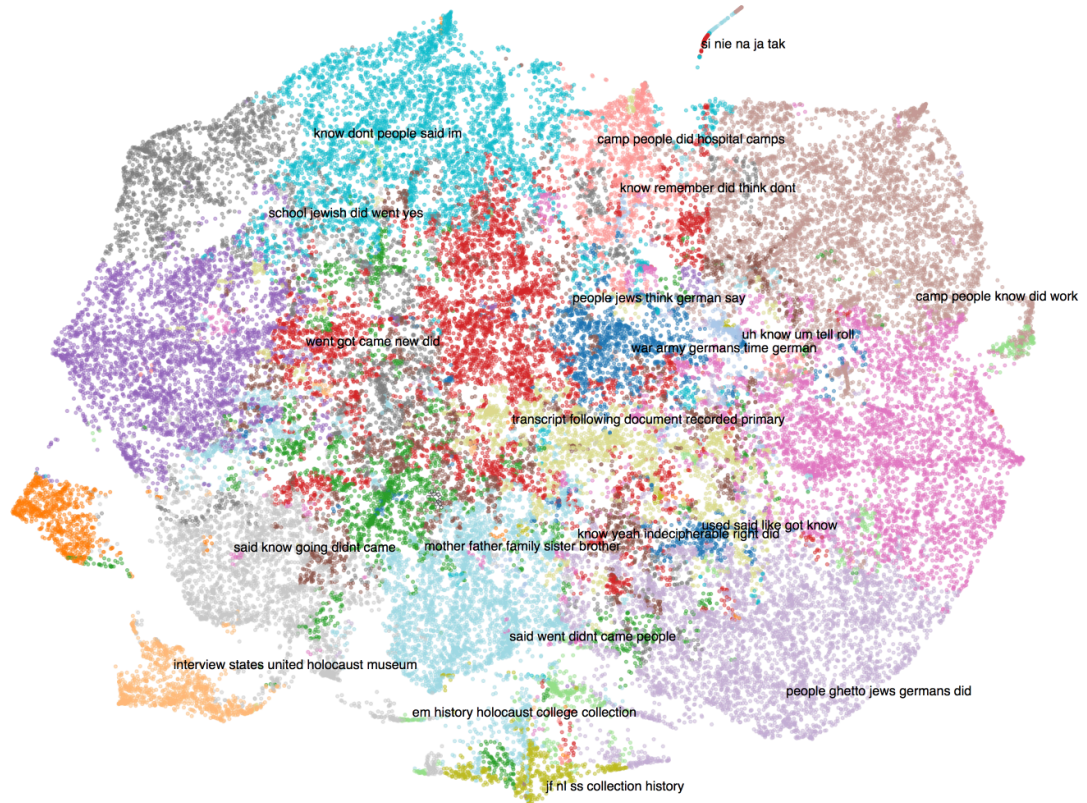
Fig. 1: Topics in the Testimonies

For the evaluation, we chose a small subset of the testimony paragraph and assigned them a (gold) topic. An automated topic modelling approach has to be able to at least partly match these topics if we compare the top automatically created topic with the manually assigned one. Unfortunately, we were not able to achieve a satisfactory performance with an accuracy higher than 70%. We either need to increase our effort to create a better gold standard topic collection or – more likely – the underlying text is especially difficult to datafy, as it is not just a historical collection with interviews going back to the 1970s using varying interview techniques, but also a fairly unsystematic collection of thoughts, typical to oral history collections. This is clearly indicated in Figure 1, where the 20 topics are color-coded and also described by its five top terms. While most topics are fairly distinct, others are clearly overlapping. For instance, we find parts of the archival (dark-brown) memory attached to all other memories. Memories in

non-English languages can also be found attached to all other memories (light-turquoise).

Conclusion

This paper has attempted to argue an age-old question in history from a new digital perspective. Creating evidence from sources by looking for connections and making links between events, people and places is key to all historical research. We have argued that even the most basic methods of generating evidence in history have changed because we are confronted with digital transformation of research and archives. This changes the way historians analyze evidence for historical change.

Before the digital transformation, critical evidence from collections required an understanding of physical archival processes and the metadata work of archivists. In the digital age, historians need to understand how computers identify archival documents. For critical scholarship, it is essential to understand at least the basics of computational approaches to integrate results from archival searches into the understanding of historical events. However, as we have seen from the case study, the basic step of evidence gathering - in itself a complex digital task - is not sufficiently reflected even in publications that are formally required to express their methodologies, such as postgraduate theses.

Besides, we suggest that humanities researchers who will work increasingly with digital evidence, focus primarily on developing or learning about the following three methodological items which will enable new historical practices. The first requirement is to address the humanistic critique of how computers identify collections and the tools we can develop to enable new critical historical work. EHRI has addressed this issue by innovating graph databases for historical research on the Holocaust. The graph database has been used successfully by the EHRI consortium for the integration of heterogeneous material from dispersed archives, and it allows for the integration of different types of data (records, thesauri, contextual information, etc.) into the same graph model. It offers us ad hoc search and identification functionalities based on graph traversal, and it scales well with greatly increased quantities of data against user traversals.

The second, and crucial, requirement is the development of online provenance which will enable new digital source criticism. Only if we develop new methods of online provenance, describing the derivation history of digital objects, can historical evidence be accounted for in the digital age. The provenance of something is its history or origins, so that the history of an event that has occurred, or a result that has been produced, can be traced back and explored. This is valuable for historical interpretation and contextualizing events, and is finally developing trust in the sources and their interpretation.

Finally, the typical process of iterations of research questions is challenged by new large-scale datasets that require computational approaches such as machine-learning and other statistical approaches. Here, nonparametric approaches dominate the digital history work. But the validation of these approaches, part of any critical computer reasoning, is problematic. Validation of computational models is, up to now, less integrated into the analysis of the past with computers. To this end, we need to enhance standard computational validation techniques with existing historical or social theories, and compare the results. This will close the circle towards a more traditional analysis of evidence, but it is very difficult to do and can only be executed in actual analyses of historical data. There are some beginnings of this kind of research, but much more work is needed.

_____

Tobias Blanke is a Reader in Social and Cultural Informatics in the Department of Digital Humanities and current Head of Department. Tobias' academic background is in philosophy and computer science. Before joining King's, he held various positions in the digital industries. His particular research expertise is big data from heterogeneous social and cultural collections as well as the social shaping of technologies. Tobias works on several international projects and committees and in particular led on the research and development work of the European Holocaust Research Infrastructure.

Conny Kristel is senior researcher at the Research Department of NIOD Institute for War, Holocaust and Genocide Studies. Conny holds a PhD in History from the University of Amsterdam. Before joining NIOD she worked as a freelance journalist and writer. She also held several part-time positions in academic research (project) management. Her particular research expertise is the historiography and literature of the Holocaust and the impact of war on society. Conny has been leading on several

(international) projects on the impact of the Second World War on European societies. Since 2010 she is Project Director of the European Holocaust Research Infrastructure.

How to quote this article:
Tobias Blanke, Conny Kristel, "*Historical Research and Evidence in the Digital Age. The Case for New Approaches in Holocaust Research,*" in Holocaust Archives and Research in the Digital Age, eds. Laura Brazzo, Reto Speck, *Quest. Issues in Contemporary Jewish History. Journal of Fondazione CDEC*, n. 13 August 2018

url: www.quest-cdecjournal.it/focus.php?id=399