Datafication in the Historical Humanities: Reconsidering Traditional Understandings of Sources and Data

International Conference and Workshop at the German Historical Institute Washington, June 2-4, 2022, co-sponsored by the GHI, the Luxembourg Centre for Contemporary and Digital History (C^{2}DH), the Chair of Digital History at Humboldt University of Berlin (HUB), the Consortium Initiative NFDI4Memory, the Roy Rosenzweig Center for History and New Media (RRCHNM), and Stanford University, Department of History. Made possible by grants from the Deutsche Forschungsgemeinschaft (DFG) and the Friends of the GHI. Conveners: Daniel Burckhardt (GHI Washington), Andreas Fickers (C^{2}DH), Zephyr Frank (Stanford University), Torsten Hiltmann (HUB), Jana Keck (GHI Washington), Mills Kelly (RRCHNM), Simone Lässig (GHI Washington), and Atiba Pertilla (GHI Washington). Participants: Cécile Armand (Aix-Marseille University); Jeremy Auguste (Aix-Marseille University);Francesco Beretta (Université de Lyon); Laura Brannan (George Mason University); Heiko Brendel (University of Passau); Megan Brett (George Mason University); Peter Bushell (University of Hawai'i at Mānoa); Kim Dresel (Arolsen Archives); Georg Fertig (University of Halle-Wittenberg); Meghan Ferriter (Library of Congress Labs); Catherine Foley (Michigan State University); Thora Hagen (University of Würzburg); Vanessa Hannesschlaeger (German Literature Archive Marbach); Walter Hawthorne (Michigan State University); Sarah Hendriks (Trinity College Dublin, The National Archives Kew); Christian Henriot (Aix-Marseille University); Katharina Hering (GHI Washington); Jörg Hörnschemeyer (GHI Rome); Dan Howlett (George Mason University); Pim Huijnen (Utrecht University); Nina Janz (Luxem-

bourg Centre for Contemporary and Digital History (C²DH)); Helena Jaskov (University of Luxemburg); Patrick Jentsch (University of Bielefeld); David Knecht (KleioLab GmbH); Emily Kuehbauch (GHI Washington); Jeremy Land (University of Helsinki); Darren Layne (The Jacobite Database of 1745); Zoe LeBlanc (University of Illinois, Urbana-Champaign); Benjamin Lee (University of Washington); Sharon Leon (Michigan State University); Yunxin Li (Stanford University); Katherine McDonough (The Alan Turing Institute); Katharina Menschick (Arolsen Archives); Katrin Moeller (University of Halle-Wittenberg); Laura Niewöhner (University of Bielefeld); Jessica Otis (George Mason University); Clemente Penna (Mecila - Maria Sibylla Merian Center Conviviality-Inequality in Latin America); Eva Pfanzelter (University of Innsbruck); Kristina Poznan (Michigan State University); Martin Schmitt (Technical University Darmstadt); Philipp Schneider (Humboldt-University of Berlin); Valentin Schneider (National Hellenic Research Foundation Athens Greece); Jennifer Serventi (National Endowment for Humanities); Alice Sheill (Michigan State University); Abigail Shell (Library of Congress Labs); Daniil Skorinkin (University of Potsdam); Daniel Stracke (University of Münster); Greta Swain (George Mason University); Alina Volynskaya (EPFL); Joëlle Weis (University of Trier); Daryle Williams (University of California, Riverside); Andreas Witt (University of Cologne & IDS Mannheim).

After a COVID-related delay, the Fifth Annual GHI Conference on Digital Humanities and Digital History took place in a new hybrid format which for the first time combined in-person sessions at the institute with keynotes, workshops, and a poster session opened up to remote participants and presenters.

After Simone Lässig's welcoming remarks, the conference started with Zoe LeBlanc's keynote presentation "Table for One: Anecdotes on the Cultures and Challenges of Data

(-fication) for Historians." Inspired by her 2017 viral tweet, which asked if you can do Digital History as an individual, she shed light on what datafication can look like for an individual scholar. Revisiting her initial question, LeBlanc stated that collaboration is often seen as a crucial factor for datafication, but it is unequally distributed and restricted to those lacking the necessary resources. She therefore emphasized that datasets for data-driven research need a new set of scholarly practices and interpretative frameworks in order to enable individual researchers without such means to do meaningful work.

Following a block of two parallel hands-on workshops, the first panel on "Merging Datasets from Different Archives" underlined the broad range of building and analyzing digital archives featuring both highly collaborative endeavors as well as work carried out by one or two researchers only. Andreas Witt presented the highly collaborative "EncycNet" which unites 22 German reference works from the early eighteenth to the early twentieth century. The goal is to use machine learning techniques in order to build a knowledge graph on this corpus. By extracting, matching and disambiguating the entries among the different works, it should be possible to analyze reflections of conventional knowledge both over time as well as between general and more specialized encyclopedias. Yunxin Li used both dynastic histories as well as the Database of Han Officials to investigate elite social networks and geographic mobility. "Network analysis tells us something old and something new": while earlier research recognized that groupings of influential scholars existed, network analysis now makes it possible to determine which scholars were influential. But she also cautioned that statistical observations more often reflect the structure of sources than an actual historical change. Jeremy Land presented his initial reflections and a roadmap on how to use the so-called "Bill of Ladings" to reconstruct, together with Werner Scheltjens, eighteenth-century colonial American

merchant networks from scratch. Since every bill was issued in three copies (merchant, captain, receiver), they were spread all over the globe and we find them already digitized in numerous archives. All three presentations underscored the potential of the digital to break through the traditional boundaries of a single book or archive. Aggregated corpora and collections are essential for identifying connections and trends that would otherwise be missed.

The first conference day concluded with the second panel on "How to Deal with Biased or Incomplete Data(sets)?" guided by Meghan Ferriter. It discussed how to write history on the basis of fragmented sources. Based on different text corpora, Martin Schmitt combined geographic, climatic, and soil data to follow the dispersal of seeds throughout the nineteenth century and discussed various data-driven possibilities to reconnect forestry practices long believed to be mostly self-contained. Philipp Schneider showed the potential of representing graphical sources like murals explicitly through data models on the basis of the Open World Assumption. The discussion also illustrated the potential and constraints of modeling gaps in datasets. Making the vague state of source material explicit is a first step in avoiding wrong conclusions, for example in social network analysis, and provides a starting point to the question how to make such gaps productive.

The next morning started with the second keynote, "What's in a Footnote? Datafication and the Consequences for Quality Control in Historical Scholarship," by Pim Huijnen. In the last couple of years, the debate on replicability, initiated in psychology and medicine a decade ago, reached publications in the Digital Humanities and led to questions about the validity of the statistical (and computational) methods used to promote far-reaching conclusions. By tracing the history of a "fact" from Dutch energy history through multiple citations, Huijnen showed that traditional humanities

scholarship is prone to become a victim of a game of Chinese Whispers, no matter how many footnotes are placed. Despite such weaknesses, he sees little sense in pushing humanities scholarship towards formulating clear hypotheses based on quantitative data, which can then be tested. Instead, there is a need to conceptualize "reproducibility" in the context of historical research, where questions on how things changed are as important as a provable hypothesis on why they changed. Rather than searching for mistakes ("maximal reproducibility"), we should focus on making the implicit more explicit ("minimal reproducibility").

Panel III presented "Case Studies for Research Data Management in the Historical Humanities." Francesco Beretta underlined the importance of improving the reuse of research data in the historical sciences, which he characterizes as good quality information with reliable provenance. To make this information truly interoperable, it needs to be linked against an authority file and modeled according to an ontology, a shared conceptualization across a certain domain of interest. David Knecht presented Geovistory, a browser-based research environment that enables users unfamiliar with the at times daunting concepts of the Semantic Web to create, enhance, and share properly structured research data. Darren Scott Layne illustrated such conceptual steps through "The Jacobite Database of 1745," which provides high quality prosopographical entries sourced from archival records. Valentin Schneider presented the scope and challenges of the German Occupation Database. In addition to properly datafying the historical records, the project aims to display the dynamics of war operations in space and time and finding adequate representations of military and paramilitary units' cruel impact on the daily life of the local population in Greece between 1941-1944/45. Working with predefined ontologies and off the shelf software promise to greatly simplify data management and reuse. But they won't provide an easy bridge across different conceptualizations, such as an

actor-centered approach versus one making heavy use of a historical geographic information system.

Panel IV, entitled "Turning Analog Into Digital Data," touched upon opportunities for transregional research. In his talk, Daniil Skorinkin pointed out that the largest prosopographical databases for twentieth-century Russia centered on the First and Second World Wars and the Stalin era repressions. They provide important overlaps (e.g. in people mentioned) for which record linkage is missing. In his project, a heterogeneous team of researchers aims to create a single data pool out of the different databases aligning as well as unifying personal data fields. Among others, the following discussion addressed the challenges of these tasks, such as the different spellings of cities and villages in the datasets. Clemente Penna and Eva Pfanzelter presented different transregional approaches. Penna introduced a database of Brazilian bills of exchange and stated, since the Brazilian bills were not much different from those in other commercial centers, the database could later incorporate legal and notarial records from other regions, such as England, the United States, Africa, and Latin America. Pfanzelter reflected on the workflow of "ReMigra", a project that studies the return migration in South Tyrol after the Second World War. An interdisciplinary team elaborated a reusable datafication workflow, which starts with digitization and organization and is characterized by the focus on data as well as on process orientation as well as critical reflections. The discussion that followed focused largely on the difference between the political and economic reasons for remigration. While this aspect of migration is still understudied, digital history shows great potential to fill this gap.

The fifth panel on "Research with the Public: Crowdsourced Datafication" chaired by Atiba Pertilla challenged potentials and limits of crowdsourced datafication, deliberating the impact on mechanisms of agency along three papers. Katrin Moeller and Georg Fertig analyzed the situation of

female employment-based datasets created by individual citizen scientists. They demonstrated an approach that allows re-use of data in a way that is different from its original context as written sources. By this, new methods of statistical data analysis reshape questions of emancipation in the nineteenth and twentieth centuries in phases of social crisis. Kim Dresel and Katharina Menschick linked the principal questions of the panel to the work of online archives cooperating directly with volunteers. At Arolsen archives, crowd-sourcing methods are implemented in order to index documents of the Shoah and minorities who had suffered Nazi persecution. Having new participants exploring sources also raised a range of ethical questions on how and by whom the metadata should be generated. The panel concluded with Abby Shelton's presentation, who gave insights into the institutional proceedings at the Library of Congress mobilizing volunteers to contribute to digital collections. This broadened the ethical discussion to the triangular relation of historians, sources and volunteers, asking which role the latter in their role as participants should take within the negotiations on the definition of meaningful content.

The third day started with the presentation and discussion of the posters, available at https://datafication.hypotheses.org/poster. The break-out rooms on Zoom provided a great environment for in-depth discussion on a one-on-one basis or within small groups. It was followed by a second session of three workshops, which again could be attended on-site or remotely.

After the lunch break, Daniel Stracke, Sharon Leon, and Peter Bushell introduced their projects, all of which represent different "Methodologies of Datafication" in panel VI. The aim of the "European Historic Towns Atlases," a long term project presented by Stracke, is an edition of cartographic and iconographic sources for the reconstruction of the historical topography of pre-modern urban spaces. Like

other presenters, Stracke pointed out the project's need to look for future ways to improve the workflow as well as the goal to set new standards. Leon talked about the project "On These Grounds: Slavery and the University" and her interdisciplinary team's work realizing an event-based ontology. The following discussion covered the ongoing debate on the reparation for descendants of enslaved people by Georgetown University and the future incorporation of the data into enslaved.org, the linked open data platform presented above. Her talk was followed by Bushell's presentation on the "War Crimes Documentation Initiative" about World War II-era war crimes committed by the Japanese in the Asia-Pacific region. He highlighted the issue of making data accessible to people who are not very familiar with datafication.

Elizabeth Murice Alexander chaired the final panel on "How to Create Sustainable Digital Projects." Vanessa Hannesschläger began by exemplifying the complications of retaining data that was formerly collected in a self-made database and not even modeled according to standardized data formats. The discussion faced the problems of evolving data standards and research methods and tried to shape concrete steps for a conversion of a rich data collection into a new digital edition. An honest insight into the pitfalls of developing and realizing data projects was provided by Joëlle Weis, who opened the floor to a discussion about how to handle discrepancy in the final stage of digital projects and how to make these experiences useful in regard to future projects. Helena Jaskov closed the panel with a demonstration of the data orchestration engine *Kiara,* an environment for supporting users in re-using tried and tested data as well as helping them to manage their own research data and its ongoing expansion with newly added metadata. Despite the prominent position of the "experimental" in the discourse on the Digital Humanities, talking publically about dead ends and failure is still a rare exception, which was greatly appreciated.

The three intensive conference and workshop days were summarized by Andreas Fickers in his final remarks. Datafication is a complex process that is deeply affecting our discipline. Datafication is about re- and co-constructing epistemic objects, and the models and technologies used to construct them intervene directly with our research. While this is still very much work in progress, we can clearly identify the research data thus created as a boundary object, interpreted in very different ways, for example, in library science, information technology or in the humanities.

Datafication means transformations, formatting, and these forms matter. Datafication, despite the rise of machine learning, is, at least in the historical humanities, still a mostly manual, laborious, often frustrating task prone to failure. But it is mostly collaborative work, which moves us away from "lonely research" and points to new ways of doing history, which is co-designed and co-produced. Labor always has a socio-economic component and leads to questions, such as specifying the return on investment and the riskiness of such endeavors. Epistemic virtues like trust, carefulness but also courage, willingness and honesty are important aspects to our digital knowledge co-production and will therefore stand at the center of the next Digital History conference planned for the spring of 2024, for the first time not at the GHI but in Luxembourg.

Daniel Burckhardt,
(GHI Washington)
Tim Feind
(Universität Wien)
Lara Raabe
(Universität Heidelberg)