מידע דעה
מדע עתון
Medaon מידעון
Magazin für jüdisches Leben in Forschung und Bildung

Miszelle

Yael Netzer, Amalia S. Levi

Legacy Catalogs as Data. The case of Quellen zur Geschichte der Juden in den Archiven der neuen Bundesländer

Introduction

Archival descriptions that exist as text, published as books or other types of reference works, are largely inaccessible to users today who might not know about or have access to these works. Even when archives provide online metadata for their collections, these are only a partial view of knowledge contained in legacy descriptions. Modeling and datafying the knowledge contained in published catalogs and other reference books makes such knowledge accessible to wider audiences and opens different ways of engagement with collections by diverse audiences.

This paper reports on a project of modeling and datafying knowledge in a legacy catalog that took place in September 2022 during the first DH Jewish Hackathon[1] organized by the Moses Mendelssohn Center for European-Jewish Studies (MMZ) and the Potsdam Network for Digital Humanities at Potsdam University. The Hackathon was focused on "*Digital Heritage and Jewish Studies*", part of the unconference series "Henriette Herz Hackathons", funded by the Alexander von Humboldt Foundation in the context of the Henriette Herz Award. The Hackathon took place during four days of communal efforts to advance Jewish Studies-related digital humanities research projects. Forty people with diverse sets of skills from both theoretical and computational backgrounds got together to collaboratively work and advance these research projects.

Our project was led by Yael Netzer as part of her Humboldt Research Fellowship. It had to do with the datafication of the reference work Quellen zur Geschichte der Juden in den Archiven der neuen Bundesländer (*Sources for Jewish History in Archives in the New Federal States of Germany*).[2] This datafication idea was developed during discussions prior to Netzer's arrival to Potsdam, inspired by the Yerusha project.[3]

Quellen zur Geschichte der Juden is a six-volume publication edited by Stefi Jersch-Wenzel and Reinhard Rürup, developed in 1992–2001, the first decade after the reunification of Germany. It is a comprehensive compilation of descriptions of collections of Jewish-related materials found in archives previously located in the German Democratic Republic (GDR). The publication was of great importance, since these records were in most cases unknown and inaccessible for people outside the GDR in the years after the Second World War. Additionally, despite the fact that some volumes were made

[1] https://www.uni-potsdam.de/de/digital-humanities/aktivitaeten/henriette-herz-hackathons/dhjewish-hackathon [20.02.2023].

[2] Members of the team led by Yael Netzer were: Aaron Christianson (UB Frankfurt), Elena Hamidy (UB Giessen), Imme Klages (Universität Mainz), Amalia S. Levi (Bonn University). See also: https://www.degruyter.com/serial/qgjab-b/html [20.02.2023].

[3] Yerusha is an online portal providing access to 12,619 (to date) descriptions of archival collections with Jewish material from approximately 690 holding institutions in Europe. See https://yerusha.eu/ [12.12.2022].

available by the publisher as eBooks (in PDF format), these are static files that did not allow users nuanced access (e.g., through facets). The aim of the Hackathon project was to overcome these limitations and provide scholars with data.

Books as data

Reference works such as library and archival catalogs, concordances, dictionaries and lexicons, as well as bibliographies have a long history. Their contents and structure reflect common conventions and systematic perceptions of their period's knowledge. Such works are already semi-structured in nature, meaning the information contained within those publications is arranged according to certain conventional patterns. In the digital era, this systematic nature of the information facilitates semi-automatic computational approaches for converting it into data.[4]

Although books are written using electronic devices and text editors, and reference books most likely rely on databases or spreadsheets administered by the author, most of the books published are not provided in a form that gives access to the book's data (for instance, as scanned files of the original books). It is even less common to intentionally produce a book in a format that is both human readable and machine actionable, (e.g. XML/TEI[5]). The benefit of this approach is the ability to derive various representations from the same source: a printed book, an eBook, a website or tables that can be queried, computationally manipulated or visualized, and used for *distant reading*.

The challenge we chose during the Hackathon was to convert given scanned files (in PDF format) of the six volumes into data. Well-structured data can be loaded into a dedicated web or content management application, it can be searched, queried and linked with other data – and in our case it can perhaps later be integrated into the Yerusha database. This challenge met our interests in archives, catalogs and the concept of *collections as data*.

Collections as Data [6] is a conceptual approach that promotes turning cultural heritage collections into data through various processes, such as making them as explicit and shareable, documented and accessible, structured, cleaned and normalized as possible, describing them with controlled vocabularies, enriching them by using techniques such as Named Entities Recognition, and linking them to other Linked Open Data resources. We chose this approach, along with fundamental principles of Digital Humanities, to transform the information given in the books into data that is both *human readable and machine actionable*.

[4] Converting legacy archival descriptions into data without interrogating the inherent power structures and values introduces various issues. We touch briefly on this in this article, and will discuss more extensively in an upcoming article.

[5] TEI, online at: https://tei-c.org/ [14.12.2022].

[6] Padilla, Thomas: "On a Collections as Data Imperative" (2017), online at: https://labs.loc.gov/static/labs/work/reports/tpadilla_OnaCollectionsasDataImperative_final.pdf [12/14/2020].

Methods

The first step was to *datafy* the information in the downloadable open access PDF files available on the publisher's site.

In his review of this six-volume resource, Nils Roemer[7] wrote: "One wonders if future archivists and historians will soon find a way to present this type of material in a more accessible and less cumbersome fashion. In the end, the indices provide only limited access that could be improved significantly if one were able to push the 'find' key, which is rapidly becoming a not merely familiar, but standard, means of tracking information." In our challenge, we worked on formatting the data of these volumes not only to be searchable by pressing the 'find' key, but to become a computational body of knowledge.

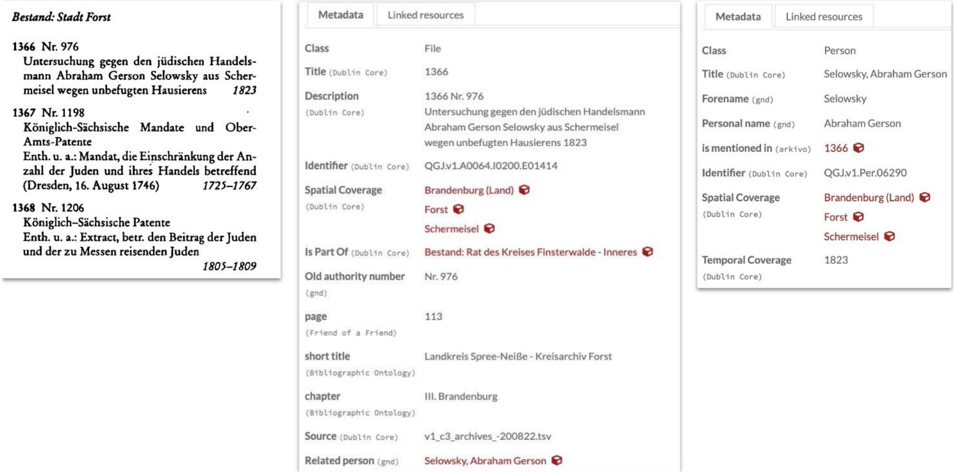
Datafication of published books is possible by using two methods that work well with reference books that are semi-structured by nature: a) decomposing the textual version into data, or b) using the original data; for example, data collected before the publication of a catalog, which ultimately are not always available. In our case, we began by applying the first method aiming to understand and re-model the information structure of the book, then extract the text and convert it into tables designed according to the model, clean the data (removing page numbers or running page headers, and identifying dates), and re-structure. For this iteration, we used regular expressions, Python code, and OpenRefine[8]. This process was applied to the main chapters with the information about the archives, the collections and the records, and to the indices at the end of the book, which consisted of entries for locations, individual persons and institutions. We linked each entry in the index to its entry in the book using unique IDs. In this way, we were able to create and query a network of archival material based on locations, organizations, and places, and to track them along time. The result is a collection of data which were subsequently uploaded as linked data to an Omeka S instance.[9]

In Figure 1, we show how the original entry is linked to its "Bestand" (collection) in the holding archives, linked to information from the index, regarding related individual persons, locations, and organizations, and with temporal coverage (based on the scale of earliest to latest year mentioned).

[7] Roemer, Nils: Review of Review of Quellen zur Geschichte der Juden in den Archiven der neuen Bundesländer, by Stefi Jersch-Wenzel and Reinhard Rürup. The Jewish Quarterly Review 92, no. 3/4 (2002): 632–632. https://doi.org/10.2307/1455472.

[8] Open Refine, online at: https://openrefine.org/ [12/14/2022].

[9] Omeka S [https://omeka.org/s/] is an open-source web-publishing platform that we used as a CMS (content management system).



Bestand: Stadt Forst

1366 Nr. 976
Untersuchung gegen den jüdischen Handelsmann Abraham Gerson Selowsky aus Schermeisel wegen unbefugten Hausierens *1823*

1367 Nr. 1198
Königlich-Sächsische Mandate und Ober-Amts-Patente
Enth. u. a.: Mandat, die Einschränkung der Anzahl der Juden und ihres Handels betreffend (Dresden, 16. August 1746) *1725–1767*

1368 Nr. 1206
Königlich–Sächsische Patente
Enth. u. a.: Extract, betr. den Beitrag der Juden und der zu Messen reisenden Juden
1805–1809

Class	File
Title (Dublin Core)	1366
Description (Dublin Core)	1366 Nr. 976 Untersuchung gegen den jüdischen Handelsmann Abraham Gerson Selowsky aus Schermeisel wegen unbefugten Hausierens 1823
Identifier (Dublin Core)	QGJ.v1.A0064.I0200.E01414
Spatial Coverage (Dublin Core)	Brandenburg (Land) Forst Schermeisel
Is Part Of (Dublin Core)	Bestand: Rat des Kreises Finsterwalde - Inneres
Old authority number (gnd)	Nr. 976
page (Friend of a Friend)	113
short title (Bibliographic Ontology)	Landkreis Spree-Neiße - Kreisarchiv Forst
chapter (Bibliographic Ontology)	III. Brandenburg
Source (Dublin Core)	v1_c3_archives_-200822.tsv
Related person (gnd)	Selowsky, Abraham Gerson

Class	Person
Title (Dublin Core)	Selowsky, Abraham Gerson
Forename (gnd)	Selowsky
Personal name (gnd)	Abraham Gerson
is mentioned in (arkivo)	1366
Identifier (Dublin Core)	QGJ.v1.Per.06290
Spatial Coverage (Dublin Core)	Brandenburg (Land) Forst Schermeisel
Temporal Coverage (Dublin Core)	1823

Figure 1: Entry from volume 1 (left); its representation in Omeka-S (middle); person in index (right).

After the Hackathon, Netzer experimented with the second method by pursuing the original database of the book from its editors. The database was provided to us in a Filemaker (V. 5) File format[10] that was converted into TSV tables.[11] The data derived from the original database is well-organized and free from mistakes originating from the OCR[12] or caused by misidentification of the structure of an entry by code or regular expressions. The data in this database is not always identical to the data in the published version because the book content was proofread and edited while preparing it for publication. Comparing the two data sources therefore results in discrepancies which might affect data quality in the original database, since mistakes have been fixed in the published text, but not in the database.

To sum, the two methods of data acquisition resulted in two sets of tables, one for each method. These tables are complementary to each other.

[10] Thanks to Daniel Burckhardt from MMZ.

[11] A simple text file formatted in a tabular structure.

[12] Optical Character Recognition: transfering an image of text into machine-encoded text.

Post-processing

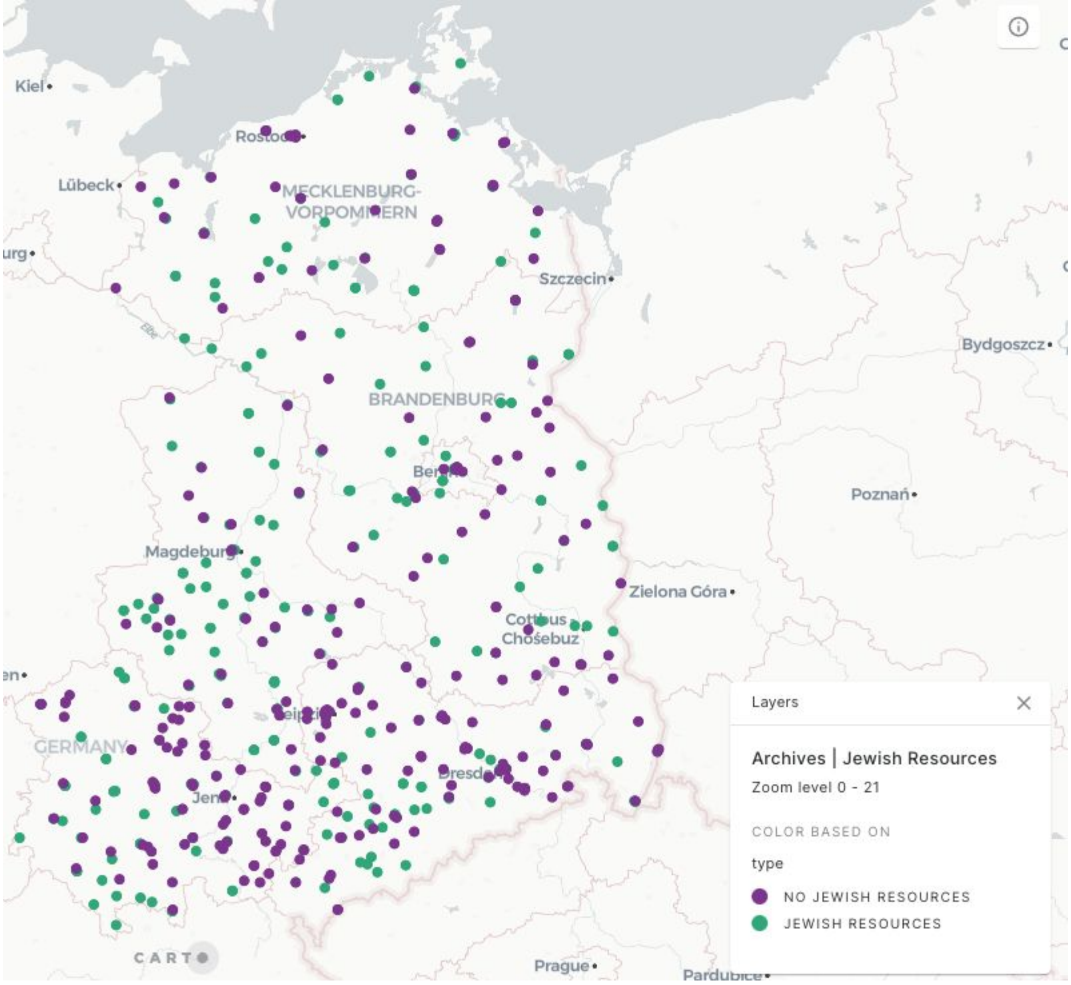


Figure 2: Map of East German archives, Volume 1, with Jewish-related records (green) or reported none (purple).

After structuring the data of the book into tables based on the model, it was possible to further manipulate the content of the dataset. For example, we enriched data on locations which were indexed with geocoding and visualized them on a map. We also geocoded the lists of archives in the book, including the list of archives with non-Jewish materials, found at the end of Volume I (Figure 2). Coordinates of the geographic locations were provided using the GeoNames[13] service.

Regarding individual persons, we searched for and enriched the data about them with additional information from Wikidata[14] or VIAF[15]. For example, we filtered individuals who lived during the 20th century and mapped their places of death when this information was available through these external resources.

[13] Geonames. Geographical Database:[https://www.geonames.org/] [12/14/2022].

[14] Wikidata: https://www.wikidata.org/wiki/Wikidata:Main_Page [12/14/2022].

[15] Virtual International Authority File (VIAF): https://viaf.org/ [12/14/2022].

Figure 3 shows that most of the 1,782 people for whom we have full information died in the Theresienstadt concentration camp.

Using such methods, the information of the printed books can be enriched with additional resources, hopefully linked to digital records at the holding archives, and can also be integrated into the Yerusha project.

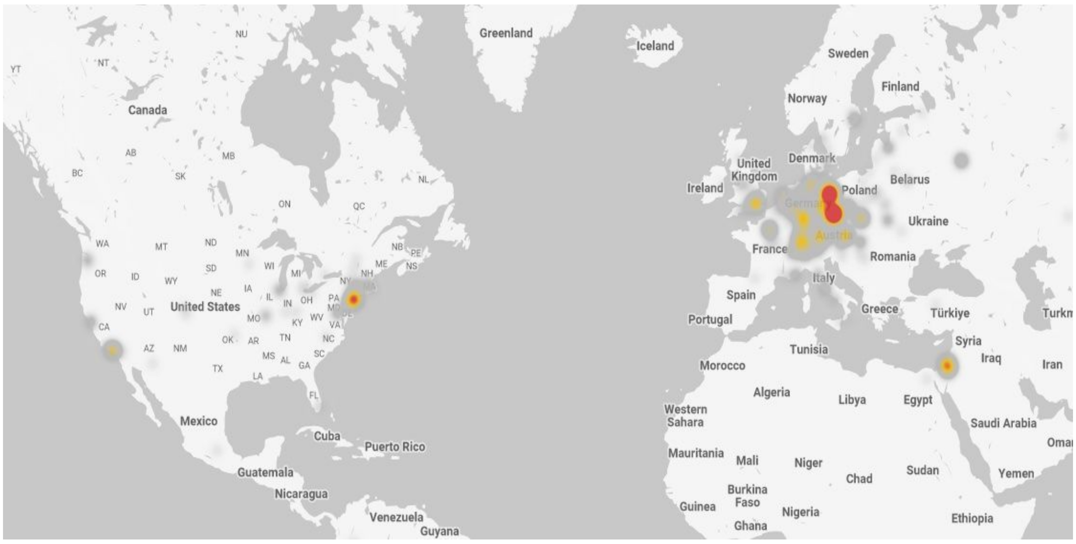


Figure 3: Place of death of persons which were found in Wikidata.

Project Results

The full text found in the *Quellen zur Geschichte der Juden* volumes is now modeled and structured in tables and uploaded into an Omeka S site. The Omeka S site and the model we used allow for simple searches through keywords, browsing the contents of the books based on their original order, and also through semantic relations (such as location, individual persons, or organizations mentioned). Search results can be downloaded, and additionally, the site can be expanded to present a timeline, quantitative analysis, possible visualizations of networks, and reconciliation with external knowledge bases.

Conclusions

We cannot overstate the importance of modeling and datafying the knowledge contained in legacy catalogs and other reference books. It makes such knowledge accessible to a wider audience and provides the tools for a variety of computational approaches. Information previously available as printed books or static PDFs can now be queried based on individual persons, time, and location or be presented as networks. It opens different ways of engagement with the material by diverse audiences.

However, the ease of access raises additional questions about the effects of digitization and datafication on knowledge, particularly when dealing with legacy descriptions that often include offensive and problematic language. It also raises questions of representation, preservation, or derivation of information. More importantly, the process of turning collections into data reveals the pitfalls of such work that, if done without contextualization, ends up reifying what is already there without interrogating underlying assumptions. As scholars doing collections-as-data work we need to be aware of such issues at every step of our work and of our role as making accessible, but also producing new knowledge.

Overall, we believe that this project contributes to Jewish Studies scholarship because it opens up information that had previously been largely inaccessible. It provides diverse users with the ability to access the content in different ways and engage with it computationally. It can also be used as a starting point to build upon, enriching it with missing information. Finally, it lends itself as a case study for reflecting over the benefits and affordances, but also the challenges and limitations of applying collections as data methods.

Citation *Yael Netzer, Amalia S. Levi: Legacy Catalogs as Data. The case of Quellen zur Geschichte der Juden in den Archiven der neuen Bundesländer, in: Medaon – Magazin für jüdisches Leben in Forschung und Bildung, 17 (2023), 32, S. 1–8, online unter http://www.medaon.de/pdf/medaon_32_netzer_levi.pdf [dd.mm.yyyy].*

About the authors:

Netzer, Yael; Hebrew University (Chief Director of Research and Teaching Fellow, Digital Humanities Center, Hebrew University; Consultant and Teaching Fellow, Digital Humanities, Haifa University; Linguist at Dicta)
Research areas: Digital and Personal Archives, Digital Humanities, Computational Linguistics, Hebrew.
Research Projects: Personal archives as autobiography; Digitization of catalogues; DraCor in Hebrew.
PhD in Computer Science, MA studies in Hebrew Literature at Ben Gurion University. Teaches Digital Humanities in Israeli Universities and the European Summer University of Leipzig.Develops and implements methods for digital personal archives, knowledge representation for archives, libraries and for the humanities.

Levi, Amalia S.; Archivist; Ph.D. Researcher (BCDSS, Bonn University)
Research areas: Archives; Jewish Studies; Slavery Studies; Digital Humanities
Current research projects: Enslaved people in Sephardic households in early modern Barbados
Main recent publications:
Zaagsma, G., Stökl Ben Ezra, D., Rürup, M., Margolis, M. and Levi, Amalia S. (2022). Jewish Studies in the Digital Age. Berlin, Boston: De Gruyter

Oldenbourg. https://doi.org/10.1515/9783110744828

Levi, Amalia S. (2019). "Intersectionality in Digital Archives: The Case Study of the Barbados Synagogue Restoration Project Archives." In B. Bordalejo and R. Risam (eds.), Intersectionality in Digital Humanities. ARC Humanities Press.