The Archive as a Graph – Provenance Research on Bundesarchiv B323

Ruth von dem Bussche und Meike Hopp

December 17, 2021

1 Project definition

Archival documents are crucial to provenance research, witnessing dispossessions and art plunder during WWII. One of the most important archival collections is the so called B323 holdings of the Treuhandverwaltung von Kulturgut (Trusting Administration of Cultural Property) in Munich, which contain important records of the Nazi authorities on the loot of cultural property within the German Reich and the occupational zones, which have been compiled primarily by the American Allies until 1949, and which are now located in the Bundesarchiv (Federal Archives) in Koblenz. Its subcollections include archival documents coming from acting parties such as the Einsatzstab Reichsleiter Rosenberg (ERR) and other German authorities or officials involved in the dispossessions as well as documents created by Allied forces, who tracked down many artworks (scattered for safeguarding in so called Central Collecting Points) for repatriation and restitution.

Although the B323 collection has already been digitized, it is not yet searchable in full text. Primary access to the voluminous files can therefore only take place via the finding aids, whose complex hierarchical structure encourages graph-based views and analyses. The finding aids follow a tree-like structure with archival sub-collections/units structured according to thematic contexts (e.g. Nazi authorities or regional focuses). But, however, by their nature, they have intervening nodes that contain information that goes beyond this strict archival tectonics. For example, the ERR photo collection is one of these subgroups, catalogued as seemingly cohesive in itself but thus having intermediate nodes related to various other archival collections (records such as correspondences, lists, index cards) that are grouped in other subunits due to historical authority structures of either the Nazis or the Allies.

Our paper outlines how this access via the basic hierarchical structure can be overhauled by a knowledge graph, including processes of machine operations, aggregation with data from other sources, matching disambiguation, and recent results of expert research to provide more comprehensive analysis and research access to documents.

Searching and navigating documents through archival platforms such as the Federal Archives’ Invenio, are still far from enabling meaningful queries across this structure, as basic research is based on text searching in fields such as title and abstract. Machine-based operations can provide an initial starting point that then requires further intellectual work. Our methodology is thus not intended to compete with the work of the archives, but to show what graph-based approaches can do to make finding aids more efficient for archival holdings such as B323, that are highly frequented by international researchers.

Our methodology goes beyond what archival finding aids can produce and does not mean to compete with the archivists’ work.

2 Data used, recognition and disambiguation

We started by parsing the EAD XML data available as open data on the archive’s website. This was transferred into Neo4J which allowed us to better understand and explore it’s hierarchical nature.

Further content analysis is based on the summary texts and archival descriptions that summarize the content of the sub-collections and include names of individuals and various organizations involved. So far, we have not taken a specific approach to including text from the documents, but rather have explored different approaches, e.g. Google NER to extracting these entities (which

---

1https://invenio.bundesarchiv.de/invenio
Figure 1: B323 graph, collections and intermediary node.

Figure 2: B323 graph.
have been recorded very inconsistently in the finding aids). Entities have been extracted and
then written back to the graph. This produces a first rough insight into the collection’s contents.
To improve the results we used different data sources for disambiguation. Wikidata is the most
obvious one. It does contain the most important names, so running Google NER plus Wikidata
disambiguation means couple of hours work and leads to a data set which gives a good basic insight
into the collection.

But for a more comprehensive understanding we need to refer to more domain specific sources,
which provide more names. Since neither national (GND) nor international (VIAF) portals offer
enough standardized data for the persons/actors relevant in this context so far, we chose two
datasets which fit well with the above mentioned archival records:

- the Lostart-website of the Deutsches Zentrum Kulturgutverluste (German Lost Art Foundation)
  which contains data related to former private art collections seized through the Nazi. 3

- b. the ERR project database which has more (sometimes overlapping) entities related to
  Belgium and France two of the main operation zones of the ERR. 4

In order to match the data more easily with that from the Bundesarchiv, and since both are
publicly available data, both sites have been scraped whereas the data has been collected in a
Wikibase.

The whole disambiguation process on the two resources used as a kind of substitute vocabulary
lead to a better data quality and help to sort the data. We can differentiate in:

- a. entities which already have a (public) identifier,
- b. “new” entities which will get an internal temporary (WikiBase) identifier,
- c. data that is not yet sufficiently identifiable.

Adding entities such as names, institutions, corporations, or status of ownership (e.g. art col-
lections) and events (e.g. exhibitions, transfers) to the graph help to identify relationships between
the information from different sub-collections and to create queries about those relationships. For
example, the plunder of larger art collections very often is a process in time, so art collections
have been seized in various raids at different times, which leads to mentions in different archival
collections. This relatedness between different subcollections can now be determined more clearly.

3 More entities into the graph

What other data might be relevant? For many art collections we know their (former) location or
address, which can be enriched with the Lostart and the ERR database. We can already see from
the ERR data published so far, that groups of artworks seized from the same collections have later
been "exploited" in convolutes. A graph based approach will allow a more precise analysis of these
circumstances.

Another level of provenance that must be considered with these archival records concerns the
origin and history of the documents and records themselves. The B323 collection is a mixed col-
lection of documents that were once produced by various Nazi German authorities in the course
of spoliation and looting, then confiscated by the Allies, then re-sorted and integrated into con-
texts of the records structure of the Central Collecting Point (CCP) in Munich under American
administration, and later transferred back into German authority structures. Also with regard
to the original paths of the provenance of the documents, the database of the Federal Archives
remains very superficial, so that we must provide more data based on broader documentation on
the history of the archival collections itselfs.

Once we have created such a graph, we might think about further extensions. We could add
more entities to our graph, add information about other collections from other archives, describe
the collections themselves in greater detail or find events (e.g. transfer of custody) that connect
nodes in the graph.

So we have the necessary documentation on archival holdings that we would like to see in our
graph, but so far provenance research has produced only few publicly linkable datasets, such as
the PDF guide to archival sources related to the ERR produced by Patricia Grimsted. 5 But even
this guide which is consulted by provenance researchers worldwide, has not yet been evaluated at
the data level.

3https://www.lostart.de/Webs/DE/Provenienz/JuedischeSammler/Index.html
4https://www.errproject.org/jeudepaume/owner_list.php
5https://www.errproject.org/guide.php and [Gri19].
4 Software solutions in the process

The process of enriching the graph is a continuous one. Machine learning like Google NER can offer a quick glimpse on the data, but more detailed analysis needs more domain specific data sets. Wikibase is a central component in this process that helps us to reconcile and match the recognized entities and to merge different identifiers for entities, e.g. a GND ID and Wikidata ID for a person. In case we do encounter entities which do not have any references in linked open data or public resources, we will be able to define new nodes and identifiers. All the reconciliation work is done with OpenRefine.

Because of its visual features we currently process the the graph in Neo4J bloom. Neo4J is also useful when it comes to aggregating data or to apply more data analysis. Wikibase could also serve as a repository for the archived data itself, which would have the advantage of allowing SPARQL queries from Wikibase services. On the other hand, the resulting dataset(s) could also be easily published as data via Github after the active work in the project is finished.

So far, we have relied entirely on descriptions from archival finding aids for provenance research and consulted them in a more or less “analog” way; there was hardly no quantitative or qualitative analysis or autopsy of the files themselves and even less extending beyond that. The latter, of course, would require much more expertise and intellectual work, but could also contribute to more efficient tools, which, thanks to easy access, could be used more intuitively, collaboratively and, above all, internationally.

5 Methodology

There have been previous efforts to explore the benefits of linked data and semantic web technology on archival collections. Documents from the period of the German occupation of the Netherlands, for example, have been dealt with in the Dutch War Collections project Oorlogsbronnen.

Another notable example from the field of provenance research is the Archivführer zur deutschen Kolonialgeschichte (Archive guide to German colonial history) which has created an important data set, cataloging archival collections and linking these to Wikidata items.

The Wikibase resource is thus essential to ensure quality doing recognition and disambiguation. Is also offers the possibility to work on these entities and include expert knowledge.

So there is an urgent need to build up our own domain related data for disambiguation and reconciliation. This does not have to be purely manual work, but could also feed on data from other providers that we prepare for enriching our own resources. Unfortunately platforms like proveana fail to give access to raw data or to provide an API and just allow simple searching on the platform itself.

So far, there is no formal ontology in our approach, but we see a great benefit in working with other related projects to develop one in the medium to long term. At the moment, however, entity recognition and graph enrichment with data - even at the informal level, e.g. via Wikibase - is more appropriate. Only where formal ontologies already exist can machine learning methods bring benefits. Projects from our still very young field of science often seem to be cut off from the discourse when applying semantic web technologies and ontologies, but we would like to change that.

How can we profit from other projects? In the Netherlands in particular, numerous LODs have already been created on World War II. While there will probably never be the ONE archival repository, resource, or database that “contains everything we need” for provenance research, we still want to continue to explore what benefits we can achieve from future collaborations with all of these projects.

References


For a summary see [Gri19];[Gra15]; [Niu16]; [LRT18]; [HGvE20]; [MTG18].

See [BHP20] and https://www.oorlogsbronnen.nl.


The problem of failed machine based recognition and disambiguation in data related to WWII art looting is addressed by [Zuc20], she also emphasizes the importance of domain experts. A different approach to integrate experts is taken by [DABS17] in a nichesourcing project.

https://www.proveana.de.

For a more ontology based approch see [Sto21].


