

German National Socialist Injustice on the Semantic Web: from Archival Records to a Knowledge Graph

Mahsa Vafaie*^{1,2}

¹*FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany*

²*Applied Informatics and Formal Description Methods (AIFB), Karlsruhe Institute of Technology (KIT), Kaiserstraße 89, 76133 Karlsruhe, Germany*

Abstract

Archival repositories contain vast amounts of historical data within unstructured textual documents, posing significant challenges for extracting coherent insights. This paper presents ongoing work towards an optimised workflow for constructing a knowledge graph from millions of archival records related to the “Wiedergutmachung” process in Germany. These records, documenting compensation and restitution efforts following World War II, offer insights into the aftermath of the National Socialist regime. The proposed workflow involves converting document images to machine-readable formats, ontology design, information extraction, and entity linking. Leveraging both traditional methods and transformer-based technologies, the workflow addresses unique challenges inherent in historical documents.

Keywords

Semantic Web, Digital Cultural Heritage, Digital Humanities, Linked Open Data, Optical Character Recognition, Information Extraction, Wiedergutmachung, Compensation for National Socialist Injustice.


1. Introduction

Since the 1990s, researchers from various domains have been increasingly engaged with providing access to information held within archival records, through online channels [1]. Archival repositories hold invaluable historical data, often in the form of unstructured textual documents that span decades. Extracting coherent insights from these records presents a formidable challenge due to the lack of standardised formats and the sheer volume of the data. Digitalisation pipelines emerge as a computational solution to this challenge, leveraging a combination of computer vision, natural language processing, information extraction, machine learning, and semantic analysis techniques. These pipelines, through sophisticated algorithms and methodologies, facilitate the transformation of archival documents into structured data points. The true power of digitalisation manifests in the construction of knowledge graphs – a dynamic framework that transforms discrete data points into interconnected nodes. Knowledge graphs play a pivotal role in bridging the gap between archival records and the Semantic Web. By interlinking the extracted information from archival records on the Semantic Web, knowledge graphs enable researchers to discern intricate relationships, uncover latent patterns, and traverse historical

Proceedings of the Doctoral Consortium at ISWC 2024, co-located with the 23rd International Semantic Web Conference (ISWC 2024)

✉ mahsa.vafaie@fiz-karlsruhe.de (M. Vafaie*)

ORCID [0000-0002-7706-8340](https://orcid.org/0000-0002-7706-8340) (M. Vafaie*)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

narratives that go beyond individual documents [2]. Furthermore, the existence of knowledge graphs as the backbones of information systems allows for inference of previously undiscovered knowledge from the given statements, and provides means for conducting exploratory search and knowledge discovery. On the other hand, enhancing accessibility of archival data leads to increased public participation and a better understanding of archival materials [3].

This paper proposes an optimised workflow for constructing a knowledge graph from millions of archival records from the “Wiedergutmachung”¹ process in Germany, for utilisation in the semantic portal “Themenportal Wiedergutmachung”². The term “Wiedergutmachung compensation records” in this paper, refers to collections of documents, records, and materials related to the process of compensation and restitution efforts that followed World War II and the fall of the National Socialist regime in Germany. Wiedergutmachung compensation records which contain an estimated amount of 100 km of archival documents, originate from the State Offices for Compensation (*Ämter für Wiedergutmachung* in German) installed by the German government in every German *Land* after the war. These collections contain a wide range of documents, including index cards, application forms, documents on legal proceedings, correspondence, testimonies, and other materials that pertain to individuals, families, and communities seeking compensation for forced labour, imprisonment, injuries, and other damages caused by National Socialist Injustice.

Wiedergutmachung compensation records serve as a historical account of the processes that took place to acknowledge and address the immense human suffering caused by the Nazi regime. They play a crucial role in documenting the complex journey of survivors and their families in seeking justice, recognition, and support for the harm they endured. They also provide insights into the legal, bureaucratic, and social challenges faced by those seeking compensation in the aftermath of such a devastating period in history. The Wiedergutmachung knowledge graph (Wiedergutmachung KG) contributes to our understanding of the impact of the totalitarian government in Germany on individuals and society and the ongoing efforts to address historical injustices. Construction of the Wiedergutmachung KG, illuminates the historical information hidden within unstructured Wiedergutmachung records, which were formerly tucked away in archival repositories, only accessible to archivists and a limited number of individuals entitled to see them for family or scientific research purposes. For instance, Wiedergutmachung KG can display connections between “claimants” and “compensation decisions”, elucidating trends and disparities within the Wiedergutmachung process.

The proposed digitalisation workflow starts with conversion of document images (i.e., scanned documents) to machine-readable formats through Optical Character Recognition (OCR). Subsequently, the workflow entails ontology design, information extraction, and linking entities with external data sources, to construct the Wiedergutmachung KG. Due to the historical nature of the documents, each of these stages present unique challenges that can be tackled using the more traditional methods, while the advancements in transformer-based technologies can also be used to address them with a more direct approach. The speed of technological advancements in the field of AI calls for a dynamic pipeline with a modular design that allows for substitution or

¹“Wiedergutmachung” is a German word that translates to “making good again” or “making amends”. In the context of National Socialism in Germany and its aftermath, it specifically refers to the efforts made to compensate survivors and victims for the losses they suffered during the rule of the Nazi regime.

²<https://www.archivportal-d.de/themenportale/wiedergutmachung>

combination of rule-based methods for accuracy with nascent AI technologies for optimisation. Keeping this consideration in mind, the same digitalisation workflow can be extended to further use cases and Digital Humanities research can benefit from the lessons learnt during the design of such a pipeline.

The remainder of this paper delves into details of the different components of the proposed digitalisation pipeline for Wiedergutmachung compensation records and discusses the intricacies of working with historical archival records. Section 2 outlines similar efforts for construction of domain-specific KGs in the context of Cultural Heritage. In Section 3 the research questions are introduced and methodologies for addressing them are discussed. Section 4 concludes the paper and sketches the next steps for future work.

2. Related Work

As archives undergo mass digitisation and the volume of digital records grows, there arises a rich but underutilised resource for researchers in the Digital Humanities. Integration of data from historical archival records into the Semantic Web and transformation of archival data according to Linked Open Data (LOD) principles has been receiving significant attention by scholars [4].

The Sampo series of semantic portals is a pioneering effort in applying Semantic Web technologies to showcase Finland's national heritage [5]. These portals utilise the modular FinnONTO, as a taxonomy of Cultural Heritage Objects [6]. WarSampo for example focuses on harmonising and publishing heterogeneous datasets related to World War II in Finland as LOD [7].

The Jewish Contemporary Documentation Center has created an online LOD database³ focused on Italian Holocaust victims and persecution events. Additionally, they've developed an associated application for utilising this valuable data [8].

In the Netherlands, "Oorlog voor de Rechter" ("War in Court")⁴ aims to unlock historical knowledge by making The Central Archives of the Special Jurisdiction (CABR) accessible online, leveraging advanced technologies and a user-centric design. CABR is the largest war archive in the Netherlands, containing files of over 400,000 people suspected of collaboration with the National Socialist regime in Germany.

The European Holocaust Research Infrastructure (EHRI) Portal⁵ serves as a valuable resource for researchers and historians interested in Holocaust-related archival material. It provides access to electronic finding aids, inventory information on institutions holding Holocaust-related records, and vocabularies related to archival descriptions [9]. Researchers can use these vocabularies to improve searchability and interoperability.

In Germany, to the best of our knowledge, this work marks the first effort to develop an LOD-based semantic portal from archival records pertaining to World War II and National Socialist Injustices.

³<http://dati.cdec.it/lod/shoah/website/html>

⁴<https://www.huygens.knaw.nl/en/projecten/war-in-court/>

⁵<https://portal.ehri-project.eu/>

3. An LODification workflow for Wiedergutmachung compensation records

Transformation of data hidden within Wiedergutmachung compensation records into LOD for increased accessibility, interoperability, explorability, and semantic enrichment is the main goal of this work. Therefore, the overarching research question in this work is: **What is the most efficient pipeline for transformation of historical archival records into a Knowledge Graph-based information system, for integration into the Semantic Web and publication as Linked Open data?** In order to accurately address the overarching research question, it seems necessary to break it down into the different components of such a pipeline, for an informed design decision. The research questions derived from a modular design for this pipeline are as follows:

- RQ1:** What advanced techniques and methodologies can be developed to improve the accuracy of text recognition for digitised archival records with challenging characteristics such as faded ink, handwritten annotations, and non-standard fonts?
- RQ2:** What are the most effective Information Extraction (IE) techniques for accurately and efficiently identifying and retrieving structured data, such as names, dates, and locations, from digitised archival records?
- RQ3:** How can existing ontologies be adapted and extended to develop an ontology for representation of archival historical records, that accurately reflects the hierarchical structure of archival records, semantic annotation of the records, and information about the agents involved in the creation and archiving of the records
- RQ4:** How can we establish reliable links between historical entities (e.g., people, places, and events) extracted from digitised archival records and relevant external databases, authority files, or reference materials?

The methodologies and initial experiments for addressing RQ1, RQ2, and RQ3 are explained below. Solutions for RQ4 are yet to be explored as a part of future work.

3.1. RQ1: Ontology Development

The development of Wiedergutmachung KG hinges on the creation of an ontology capable of modelling relationships among archival documents, court proceedings, individuals, and organisations involved in the compensation application, decision-making, and document archiving. Ensuring the validity and reliability of this model requires incorporation of domain experts' requirements and knowledge. Archivists from the State Archives of Baden-Württemberg ⁶ collaborated with the author to formulate a list of competency questions ⁷, serving as the foundation for the ontology's conceptual modeling. These questions, catering to researcher-s/historians and relatives/dependents of persecuted persons, provide insights into the domain's scope, structure, and concepts.

⁶<https://www.landesarchiv-bw.de/>

⁷The full list of competency questions is published on GitHub, in the Wiedergutmachung repository.

According to the competency questions and based on the best practices in the field of Ontology Design, the CourtDocs Ontology [10] is created to consist of three main building blocks, reusing existing ontologies in order to avoid redundancy and thereby to also enable interoperability with external data sources. Each of these building blocks and the ontologies that have been reused for their creation are described below.

Archival Hierarchy and Provenance. The Records in Context ontology (RiC-O⁸) [11] is employed to model the hierarchical structure of Wiedergutmachung compensation records, due to its inclusion of named individuals, which makes it adaptable across institutes with different archival systems and practices. On the other hand, RiC-O's incorporation of smaller entities improves findability and enables a more detailed representation of archival resources, including constituent parts like stamps [12].

Court Procedures. The PROV Ontology (PROV-O⁹) [13] is reused to depict the Wiedergutmachung process within the court system. It offers a standardised approach for modelling how entities and activities evolve over time, making it effective for process modelling. Additionally, PROV-O is widely recognised for representing provenance information, making it suitable for capturing the relationships between Wiedergutmachung procedures and the records generated or utilised at each stage of the process.

Biographical Information of Persons Involved. CIDOC-CRM is employed as the ontological foundation for representing the biographical information of individuals involved in the compensation process¹⁰. The use of a harmonising data model facilitates connection with other materials and external sources. Moreover, the event-centric approach of CIDOC-CRM, in which an individual's existence is perceived as a series of interconnected events spanning across time and space [14], enables the representation of crucial life events in prosopographical research on victims of National Socialism, including events like deportation and imprisonment.

3.2. RQ2: OCR Quality Enhancement

Creation of transcripts from scanned documents using OCR systems, greatly accelerates and streamlines the retrieval of information. However, OCR system efficacy is contingent upon factors such as text and font styles. OCR systems are typically specialised for either machine-printed or handwritten text due to their distinct visual characteristics [15]. Yet, archival documents often feature mixed text. Traditionally, workflows dealing with a variety of text types on scanned documents employed distinct text recognition models. In [16] and [17] we propose a pipeline for separation of machine-printed text and handwritten text on historical archival documents that contain both text types. This OCR pre-processing step helps improve the quality of the transcripts, by breaking down each document image into two layers, each containing a particular text type, namely, handwritten text, or machine-printed text, and consequently, feeding the layers into the appropriate OCR or Handwritten Text Recognition (HTR) engines. In our preliminary work, we achieved an increase of 16% compared to the baseline systems for separation of text types, with models trained on modern documents. In a more recent development, the new Transformer-based OCR (TrOCR) models have demonstrated the capability to

⁸<https://www.ica.org/standards/RiC/ontology>

⁹<https://www.w3.org/TR/prov-o/>

¹⁰<https://cidoc-crm.org/>

adapt to variations in fonts, text types, styles, and languages [18, 19], skipping the pre-requisite steps of dataset synthesis and model training for text type separation. With word accuracy as an OCR evaluation metric ranging between 75% and 85%, TrOCR engines from Transkribus¹¹ have the potential to optimise the OCR quality improvement step. A qualitative evaluation of the text-type separated transcripts is yet to be done, for comparison with TrOCR results.

3.3. RQ3: Information Extraction

Wiedergutmachung compensation records constitute of different document types, such as application forms and index cards, and different layouts for each document type that vary based on time and across different institutes. In a traditional information extraction pipeline, this necessitates implementation of an automatic document identification system that classifies the documents based on their type or layout, and feeds them into the respective information extraction script customised for each document type. Our experiments on rule-based information extraction using Apache Uima Ruta [20] with a subset of 75 documents with three different layouts show an accuracy of 75%, for exact matches only. However, with the advance of Large Language Models (LLMs), there is an opportunity to streamline the information extraction process, instead of laboriously crafting multiple scripts for each document and layout type. In the proposed LLM-based approach, a unified prompt, coupled with the appropriate context, can facilitate the extraction of information from all the documents spanning across document types and layouts. The quality of information extraction from archival records with LLMs is still to be evaluated for a more accurate analysis and comparison with the rule-based methods.

4. Conclusion and Future Work

The presented research provides a significant contribution to Digital Humanities research, particularly on topics related to World War II and National Socialist Injustice. Furthermore, the findings from implementations of different techniques on archival records can be extended to similar efforts on transformation of archival records into LOD.

In the next stages of the research, the focus will shift towards implementation of transformer-based technologies in the LODification pipeline and comparing the performance of these methods against the more traditional methods for each of the constituent pipeline components. Moreover, RQ4 from Section 3 will be addressed to interconnect the KG with external sources and authority files. This is a crucial step to facilitate content-based and federated semantic search and to enrich the KG. In case of the Wiedergutmachung KG, apart from Wikidata¹², there are several other knowledge- and databases, representing data on German figures and the victims of National Socialism that can be interlinked with the KG. It is also necessary to map all the extracted information to specific unique entities (e.g., persons) by means of disambiguation and entity resolution techniques.

Acknowledgments

This work is funded by the German Federal Ministry of Finance (*Bundesministerium der Finanzen*)

¹¹<https://www.transkribus.org/de>

¹²<https://www.wikidata.org/>

and supervised by Prof. Dr. Harald Sack.

References

- [1] W. Duff, Archival mediation, *Currents of archival thinking* (2010) 115–136.
- [2] J. Waitelonis, H. Sack, Towards exploratory video search using linked data, *Multimedia Tools and Applications* 59 (2012) 645–672.
- [3] J. Oomen, M. van Erp, L. Baltussen, Sharing cultural heritage the linked open data way: why you should sign up, in: *Museums and the Web 2012*, 2012.
- [4] A. Hawkins, Archives, linked data and the digital humanities: increasing access to digitised and born-digital archives via the semantic web, *Archival Science* 22 (2022) 319–344.
- [5] E. Hyvönen, Digital humanities on the semantic web: Sampo model and portal series, *Semantic Web* 14 (2023) 729–744.
- [6] E. Hyvönen, K. Viljanen, J. Tuominen, K. Seppälä, Building a national semantic web ontology and ontology service infrastructure—the finnonto approach, in: *The Semantic Web: Research and Applications: 5th European Semantic Web Conference, ESWC 2008, Tenerife, Canary Islands, Spain, June 1-5, 2008 Proceedings* 5, Springer, 2008, pp. 95–109.
- [7] M. Koho, E. Ikkala, P. Leskinen, M. Tamper, J. Tuominen, E. Hyvönen, Warsampo knowledge graph: Finland in the second world war as linked open data, *Semantic Web – Interoperability, Usability, Applicability* 12 (2021) 265–278. URL: <https://doi.org/10.3233/SW-200392>. doi:10.3233/SW-200392.
- [8] R. Sprugnoli, G. Moretti, S. Tonelli, et al., Lod navigator: tracing movements of italian shoah victims, *Umanistica Digitale* (2019) N–A.
- [9] T. Blanke, M. Bryant, M. Frankl, C. Kristel, R. Speck, V. V. Daelen, R. V. Horik, The european holocaust research infrastructure portal, *Journal on Computing and Cultural Heritage (JOCCH)* 10 (2017) 1–18.
- [10] M. Vafaie, O. Bruns, N. Pilz, J. Waitelonis, H. Sack, CourtDocs Ontology: Towards a Data Model for Representation of Historical Court Proceedings, in: *Proc. of the 12th Knowledge Capture Conference 2023*, 2023, pp. 175–179.
- [11] F. Clavaud, T. Wildi, ICA records in contexts-ontology (RiC-O): a semantic framework for describing archival resources, in: *Proc. of Linked Archives Int. Workshop 2021*, 2021, pp. 79–92.
- [12] M. Vafaie, O. Bruns, N. Pilz, D. Dessí, H. Sack, Modelling Archival Hierarchies in Practice: Key Aspects and Lessons Learned, in: *6th Intl. Workshop on Computational History (HistoInformatics 2021)*, Online event, September 30–October 1, 2021, volume 2981, Aachen, Germany: RWTH Aachen, 2021, p. 6.
- [13] T. Lebo, S. Sahoo, et al., PROV-O: The PROV ontology, *W3C recommendation* 30 (2013).
- [14] J. A. Tuominen, E. A. Hyvönen, P. Leskinen, Bio CRM: A data model for representing biographical data for prosopographical research, in: *Proc. of the 2nd Conf. on Biographical Data in a Digital World 2017 (BD2017)*, CEUR Workshop Proceedings, 2018.
- [15] N. Islam, Z. Islam, N. Noor, A survey on optical character recognition system, *arXiv preprint arXiv:1710.05703* (2017).
- [16] M. Vafaie, O. Bruns, N. Pilz, J. Waitelonis, H. Sack, Handwritten and printed text identifi-

- cation in historical archival documents, in: Archiving Conference, volume 19, Society for Imaging Science and Technology, 2022, pp. 15–20.
- [17] M. Vafaie, J. Waitelonis, H. Sack, Improvements in Handwritten and Printed Text Separation in Historical Archival Documents, in: Archiving Conference, volume 20, Society for Imaging Science and Technology, 2023, pp. 36–41.
 - [18] M. Li, T. Lv, et al., Trocr: Transformer-based optical character recognition with pre-trained models, in: Proc. of the AAAI Conf. on Artificial Intelligence, volume 37, 2023, pp. 13094–13102.
 - [19] P. B. Ströbel, T. Hodel, W. Boente, M. Volk, The Adaptability of a Transformer-Based OCR Model for Historical Documents, in: Intl. Conf. on Document Analysis and Recognition, Springer, 2023, pp. 34–48.
 - [20] P. Kluegl, M. Toepfer, P.-D. Beck, G. Fette, F. Puppe, Uima ruta: Rapid development of rule-based information extraction applications, Natural Language Engineering 22 (2016) 1–40.