Call for Master Thesis

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

Quality assured information extraction from historical address books for knowledge graph construction

Objective of this work:

Cultural heritage institutions store and digitize large amounts of data inside archives to make archival records findable and accessible for archivists, scientists, and the general public. The important heritage value for both socio-historical work as well as family and genealogical research present person registers, which contain personal information, e.g. family relations, names, occupations, places of residence, birth or death date, etc. Knowledge graphs can be used to interconnect these historical data also with external data sources and make them findable and accessible for research and education purposes as well as for the general public.

In this work, the challenges of structuring heterogeneous historical data and integrating it into a knowledge graph will be addressed. The historical sources investigated in this work are German language address books with name, address, profession and other records of persons, who resided in Nuremberg in 1910. The resources are already available in a raw textual digital form (OCR). The data obtained from the address books has to be transformed to semi-structured data using existing Natural Language Processing tools for information extraction, error correction, and normalization.

The quality of the achieved results has to be evaluated via comparison with a previously created ground truth. The identified entities as well as their relations and classes will be integrated into the knowledge graph for exploration purposes. Moreover, the entities will be linked to external sources, as e.g. Wikidata [1] and GND [2].

Task Description:

1. Segmentation and normalization of transcripts of historical address books using automated and semi-automated postcorrection methods.
2. Data normalization via resolution of abbreviations, acronyms and initials via existing dictionaries and reference lists.
3. Compilation of ground truth transcriptions for evaluation and validation.
4. Integration of the data into the Nürnberg knowledge graph.
5. Data enrichment via internal and external mapping.
6. Provision of the developed software in a reusable, well documented, and easy adjustable form.

[1] https://www.wikidata.org/
[2] https://www.dnb.de/EN/Professionell/Standardisierung/GND/gnd_node.html

The master thesis will be supervised by **Prof. Dr. Harald Sack, Tabea Tietz and Oleksandra Bruns**, **Information Service Engineering at Institute AIFB, KIT, in collaboration with FIZ Karlsruhe.**

Keywords:
Knowledge Graphs, Cultural Heritage.

Pre-requisites:
Knowledge of Programming with Python.

Contact persons:
Tabea Tietz
tabea.tietz@kit.edu
Oleksandra Bruns
oleksandra.bruns@kit.edu

aifb
Institute of
Applied Informatics and
Formal Description Methods
http://www.aifb.kit.edu/

Karlsruhe Institute of Technology