Human-Centered Computing
RESEARCH GROUP

Freie Universität Berlin

Guided ontology-backed semantic annotation with neonion

Bachelors thesis: Using Wikidata as a referential Knowledge Organisation System for entity linking and vocabulary alignment

Jakob Höper
hoeper@zedat.fu-berlin.de
Institute of Computer Science
Freie Universität Berlin

Thesis Advisor:
Prof. Dr. Claudia Müller-Birn
clmb@inf.fu-berlin.de
Human-Centered Computing
Institute of Computer Science
Freie Universität Berlin

neonion is a light-weight collaborative workbench for scholarly reading and annotation of documents [1]. Apart from basic forms of annotation, like highlighting and commenting selected portions of text, the software provides a model for semantic information extraction tasks like concept tagging and annotating relations [2]. It also contains modules for automation of such tasks, and even comes with a triple store engine for retrieval of semantic statements.

Although its technical capabilities and conceptual complexity are no unique features among the various existing approaches toward semantic annotation and enhancement, neonion aims at optimizing usability and follows paradigms for synergystic interaction between users and underlying automated processes. This focus on usability and mixed-initiative user interfaces [5] is a consequence of the observation that researchers who don't come from a technical background won't adopt semantic technologies as commonly as their collegues in life sciences or technical domains. While the latter seem eager to build extensive and interlinked ontological resources for coverage of knowledge domains from biology or medicine in order to enhance access to their publications and reception of their research, scholars from the humanities who could make valuable contributions as subject-matter experts on their respective field of research often hesitate to use available tools, expecting them not to be designed for their needs or to require too much prior knowledge about semantic web technologies.

Development of features and extensions of the neonion semantic annotation software perpetually aims at meeting the needs of subject-matter experts who are interested in generating explicit semantic data from their research and publications, but might be discouraged by tooling with perceivedly less user-friendly design. In order to offer usable tooling while at the same time enabling researchers, authors and editors to generate structured extensions to their content in a way that yields interoperable and schematically sound semantics, we propose a semi-automated approach which interacts with the user based on their input and lets them choose from recommendations for vocabulary alignment.

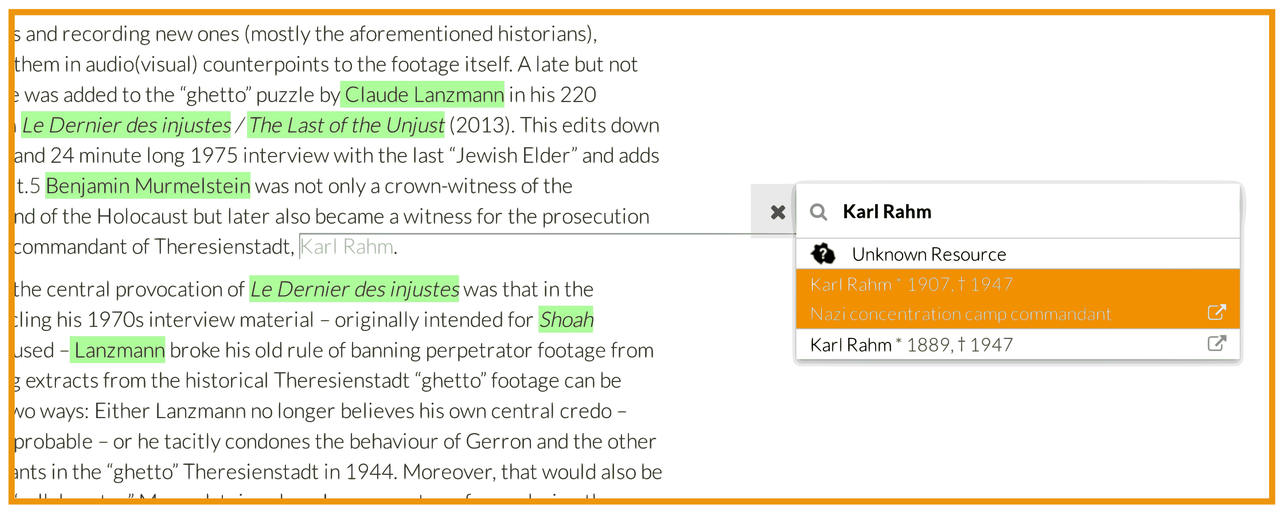
Free collaborative knowledge base Wikidata serves as a semantic backbone for structured data in Wikimedia projects. It is meant to cover of all public knowledge rather than selected scholarly and scientific domains and its large community of currently almost 20 000 active users and workflows for import of other knowledge bases such as Freebase seem to promise continuing progress. neonion is genuinely designed to allow connection to external knowledge bases in its annotation model and the internal knowledge organisation system it uses as a vocabulary for annotation. However, designing an ontology to formally define terminology and semantics for a specific research field requires not only subject-matter knowledge and labour time, but also comprehension of and skill in employed technologies.

A closer connection of neonion and Wikidata is expected to archieve mutual benefits.

- Use terminological knowledge gathered from Wikidata to extend annotation vocabulary: Align vocabulary with domain of annotated content.

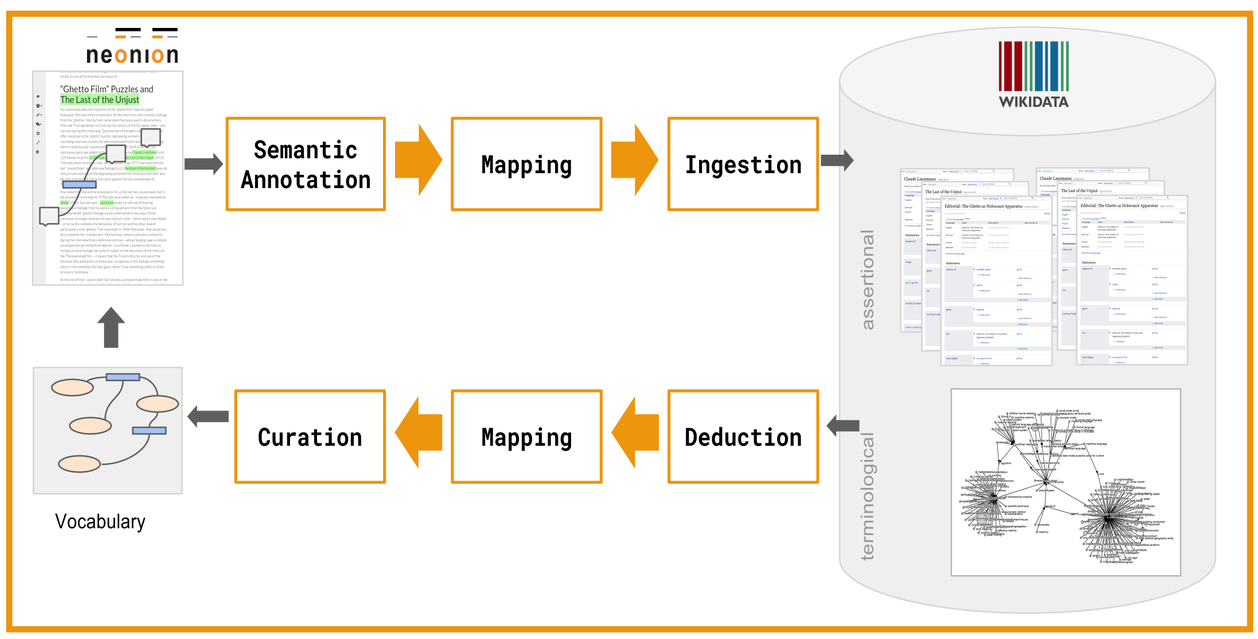
- Use mapping of vocabulary elements to extract Wikidata statements from neonion annotations: Add new factual claims to Wikidata or confirm facts known to Wikidta with bibliographic references to annotated publication.

neonion's user interface for **annotating relations**. Occurences of Named Entities are classified by annotations assigning concepts from a controlled vocabulary. Available properties for description of relationships are visually indicated as hinted connections between entities of applicable class membership.



Entity linking in neonion, using Wikidata item page identifiers for contextualization. When users select some text in annotation mode, a search dialog appears next to the document and Wikidata is queried for matching items. Results are restricted to matching instances of appropriate classes in order to ensure valid input.

The proposed functionality is being added to neonion as a propotypical implementation targeting a specific workflow involving scholarly texts from an scientific online journal. In order to assess whether Wikidata as a community repository for common knowledge can be used to extract domain ontologies based on subjects covered in scholarly texts, the **open access online journal Apparatus** [3] was chosen as a specialized textual resource, with cinematographic and historical studies with focus on central and eastern Europe as the specific knowledge domain to be covered. In order to generate candidates for vocabulary amendments, association rule mining property recommender system Snoopy is used [4].



Proposed Workflow of mutually benefitial employ of Wikidata as an authorative ontological reference. Users with specialized research contents and subject-matter expertise use a basic vocabulary for annotation of structured information from texts (upper left corner). A mapping from neonion's knowledge organisation system to Wikidata allows for optional contribution of extracted statements to the Wikidata knowledge base directly from within the neonion workbench. On the other hand, existing descriptions of identified entities are retrieved from Wikidata and schematic knowledge is deduced (bottom right corner). In combination with an association rule mining-based property recommender, neonion is able to suggest candidates for vocabulary extension to the user, who can thereby align new annotation choices with terminological semantics based on the actual content they work with.

References

[1] Müller-Birn, Claudia and Klüwer, Tina and Breitenfeld, André and Schlegel, Alexa and Benedix, Lukas. neonion – combining human and machine intelligence. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*, pages 223–226, 2015.

[2] André Breitenfeld, Maximilian Mackeprang, Ming-Tung Hong, and Claudia Müller-Birn. Enabling structured data generation by nontechnical experts. In Manuel Burghardt, Raphael Wimmer, Christian Wolff, and Christa Womser-Hacker, editors, *Mensch und Computer 2017 - Tagungsband*, pages 181–192, Regensburg, 2017. Gesellschaft für Informatik e.V.

[4] Wolfgang Gassler, Eva Zangerle, and Günther Specht. Guided curation of semistructured data in collaboratively-built knowledge bases. *Future Generation Computer Systems*, 31:111–119, February 2014.

[3] Natascha Drubek, Anke Hennig, and Irina Sandomirskaja. Apparatus: Zur Einführung. In *Apparatus. Film, Media and Digital Cultures in Central and Eastern Europe* 1, 2015. http://www.apparatusjournal.net

[5] Eric Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 159–166. ACM, 1999.

www.neonion.org