LODE: Linking Digital Humanities Content to the Web of Data

Timo Sztyler
Data and Web Science Group
University of Mannheim

Jakob Huber
Data and Web Science Group
University of Mannheim

Jan Noessner
Data and Web Science Group
University of Mannheim

Jaimie Murdock
School of Informatics and Computing & Program in Cognitive Science
Indiana University

Colin Allen
Department of History and Philosophy of Science & Program in Cognitive Science
Indiana University

Mathias Niepert
Department of Computer Science and Engineering
University of Washington

ABSTRACT

Numerous digital libraries projects maintain their data collections in the form of text, images, and metadata. While data may be stored in many formats, from plain text to XML to relational databases, the use of the resource description framework (RDF) as a standardized representation has gained considerable traction during the last five years. Almost every digital humanities meeting has at least one session concerned with the topic of digital humanities, RDF, and linked data, including JCDL.

While most existing work in linked data has focused on improving algorithms for entity matching, the aim of our Linked Open Data Enhancer LODE is to work "out of the box", enabling their use by humanities scholars, computer scientists, librarians, and information scientists alike.

With LODE we enable non-technical users to enrich a local RDF repository with high-quality data from the Linked Open Data cloud. LODE links and enhances the local RDF repository without reducing the quality of the data. In particular, we support the user in the enhancement and linking process by providing intuitive user-interfaces and by suggesting high quality linking candidates using state of the art matching algorithms. We hope that the LODE framework will be useful to digital humanities scholars complementing other digital humanities tools. [1]

1. THE LODE FRAMEWORK

The linked open data enhancer (LODE)[2] framework is a set of integrated tools that allow digital humanists, librarians, and information scientists to connect their data collections to the linked open data cloud. The initial step is to model the respective collection with some RDF serialization. For this task, tools from e.g. the Digitized Manuscripts to Europeana (DM2E)[3] project, whose aim is to parse manuscripts and make their data available in Europeana [1], can be used. Once an RDF representation exists, the LODE framework loads the RDF representation of the collection and provides several components for browsing, integrating, and enriching the collections. In the following, we describe the three modules of LODE in more detail.

1.1 Content Browsing

The content browser allows to explore the RDF dataset in an intuitive way by providing a search-based interface that resembles those of standard search engines. Users can enter keywords to search for entities in the locally stored RDF serialization of the project content. All the objects in the RDF dataset that match a given keyword query are categorized according to their types. The search field features auto completion and allows filtering by type. The syntax for this latter filter technique is adapted from the typical search engine syntax which allows searching for terms within a specific site with the command *site:url searchterm*. In LODE the user can search with e.g. *concept:human Wittgens* for an individual whose label matches the string *Wittgens* and which is an instance of the type *Human*. In addition, the LODE search interfaces provide dynamic faceted search. The results are clustered according to the *sameAs* relations so that every unique entity is only displayed once in the result. The content browser is the starting point for navigation to the following components.

1.2 Content Linking

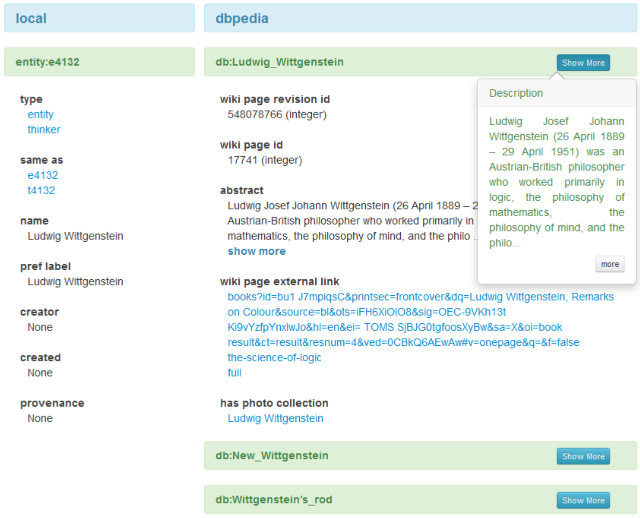
For the particular applications LODE is aimed at, the user requires full control over the linking process in order to assure the quality standard of the local RDF repository. Thus, semi-automated linking tools like SILK [2] or fully automated algorithms [3] can not be used. Hence, the purpose of the linking component is to recommend high quality suggestions from which the user then can select the correct one.

[1] This paper is a short version of the paper available at `http://arxiv.org/abs/1406.0216`.

[2] http://lode.informatik.uni-mannheim.de

[3] http://dm2e.eu/

978-1-4799-5569-5/14/$31.00 ©2014 IEEE.



entity:e4132

type
entity
thinker

same as
e4132
t4132

name
Ludwig Wittgenstein

pref label
Ludwig Wittgenstein

creator
None

created
None

provenance
None

db:Ludwig_Wittgenstein

wiki page revision id
548078766 (integer)

wiki page id
17741 (integer)

abstract
Ludwig Josef Johann Wittgenstein (26 April 1889 – 2
Austrian-British philosopher who worked primarily in
mathematics, the philosophy of mind, and the philo ..
show more

Ludwig Josef Johann Wittgenstein (26 April 1889 – 29 April 1951) was an Austrian-British philosopher who worked primarily in logic, the philosophy of mathematics, the philosophy of mind, and the philo...

wiki page external link
books?id=bu1 J7mpiqsC&printsec=frontcover&dq=Ludwig Wittgenstein, Remarks on Colour&source=bl&ots=iFH6XiOlO8&sig=OEC-9VKh13t Ki9vYzfpYnxlwJo&hl=en&ei= TOMS SjBJG0tgfoosXyBw&sa=X&oi=book result&ct=result&resnum=4&ved=0CBkQ6AEwAw#v=onepage&q=&f=false
the-science-of-logic
full

has photo collection
Ludwig Wittgenstein

db:New_Wittgenstein

db:Wittgenstein's_rod

Figure 1: Screenshot of the linking interface listing the linking candidates for the InPhO entity "Ludwig Wittgenstein".

At this point, the LODE framework supports linking to DBPEDIA [4][4] which is one of the central hubs in the link open data cloud. In addition to **sameAs** links, LODE supports different types of links modeling relationships between individuals, typed (concept), and properties. We utilize as subsets of SKOS [5] and also include several relations from the Web Ontology Language (OWL 2)[5].

The content linker performs the following steps to retrieve and display the linking candidates for a candidate entity E to the user. First, the linker component extracts a set of search terms from property assertions of entity E in the local RDF repository. To identify these terms, the algorithm maintains an extendable list of the most frequent properties describing the instance (like e.g. the label).

With the previously extracted search terms as input, the linking component generates a list of potential linking candidates for E based on two algorithms. Both algorithms are required to be interactive, returning a result ranking within seconds. Due to common hashing and indexing techniques our algorithms' complexity is sublinear with respect to the total number of possible instances. The first linking algorithm uses SPARQL queries to search for matching candidates in the LOD cloud while the second algorithm is based on the idea of exploiting Wikipedia's link structure to compute, for a given search string, the conditional probability of a Wikipedia article given the search string.

Finally, LODE extracts context for each linking candidate to help the user identify the correct alignment without overwhelming her with too much information. The context is extracted so as to help the user discriminate between entities with identical labels and names. The underlying selection process is based on the attributes frequently used across the RDF repository. Figure 1 shows a screen-shot of the linker interface.

1.3 Content Enhancing

After a **sameAs** link has been established, the enhancing module allows its users to add content from the Linked Open Data cloud to the local repository. To the best of our knowledge LODE is the first tool which supports high-quality enhancement of facts.

At a first glimpse, the reader might think that this information is already available since the **sameAs** link allows to query all available information of this entity from DBPEDIA. However, some data in DBPEDIA and other Linked Open Data sources is acquired with automated techniques and might contain inaccurate or wrong information. Since we want to perceive the high quality of the local dataset at any time, the enhancement process is essentially a verification process, in which a human domain expert verifies the correctness of facts by incorporating them in his local dataset.

The main objective of the module is to support non-technical users with an intuitive interface and high quality enhancement suggestions. In our interface the local RDF repository is displayed on the left side while DBPEDIA is located on the right side. It avoids overwhelming the user with too many potential enhancement candidates by presenting only excerpts of the most frequent class and property assertions. If the user has decided to enhance a specific class or property value, he can simply drag and drop it to the desired position. During the drag process, the user gets all possible drop areas highlighted.

Internally, LODE creates new RDF triples in the local RDF repository for each enhancement operation. In our example, LODE will add the new triple "thinker:t4132 rdfs:label 'Ludwig Wittgenstein'@en" to the local RDF repository. By keeping the target URI of DBPEDIA unchanged, it is easily possible to identify the source of the enhancement later and, thus, it facilitates provenance tracking.

References

[1] B. Haslhofer and A. Isaac, "Data. europeana. eu: the europeana linked open data pilot," in *International Conference on Dublin Core and Metadata Applications*, 2011, pp. 94–104.

[2] R. Isele, A. Jentzsch, and C. Bizer, "Silk server-adding missing links while consuming linked data.," in *Consuming Linked Data (COLD)*, 2010.

[3] J. L. Aguirre, B. C. Grau, K. Eckert, J. Euzenat, A. Ferrara, R. W. van Hague, L. Hollink, E. Jiménez-Ruiz, C. Meilicke, A. Nikolov, *et al.*, "Results of the ontology alignment evaluation initiative 2012," in *Ontology Matching Workshop*, 2012, pp. 73–115.

[4] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, "Dbpedia-a crystallization point for the web of data," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 7, no. 3, pp. 154–165, 2009.

[5] T. Baker, S. Bechhofer, A. Isaac, A. Miles, G. Schreiber, and E. Summers, "Key choices in the design of simple knowledge organization system (SKOS)," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 20, pp. 35–49, 2013.

[4] http://dbpedia.org/About
[5] http://www.w3.org/TR/owl2-overview/