Database of historical places, persons and lemmas

Natalia Korchagina [1, 2]

[1] *Schweizerische Rechtsquellenstiftung / Zürich, Switzerland*
[2] *Institut für Computerlinguistik / Universität Zürich, Switzerland*

Proposed session - **Poster session**

Keywords - **digital humanities, database, RDF triple store, multilinguality, NLP for historical texts**

The importance of the representation of humanities material as structured, interconnected objects has grown with the recent emergence of the ideas of Linked Data and Semantic Web. Efficient cataloging and storing of humanities data can facilitate the research and knowledge exchange within the field. In this respect, the use of modern technologies of data storage, i.e. databases, is a crucial point for a digital humanities project.

The Swiss Law Sources Foundation has been handling the critical edition and publishing of Swiss historical legal manuscripts for over a hundred years. By today, over 100 volumes of texts have been published, about 30 of them are available as digital editions. This collection contains texts in German, French, Italian, Romansh and Latin languages. The texts' creation time ranges from the 10th to the 18th centuries representing, for this reason, a rich source not only of historical, but also of linguistic information on language evolution. Each of these volumes contains a back-of-the-book index of persons, places and lemmas mentioned.

The creation of the database should facilitate the edition of the index of future volumes, as well as to be a starting point for users looking for information on a specific personality/place which could have been mentioned in several Foundation's volumes. The database will have the four CRUD (create, read, update, delete) basic functions of persistent storage, and will provide its users with an intuitive GUI. This database is intended for use in two directions: first, for the edition of indexes of the upcoming volumes where editors will update/create database entries via GUI instead of working with Excel files; and second, for large public browsing the database via GUI in read-only mode. To give some more details on a historical person/place the database entries will be linked (when possible) with the corresponding GND (Integrated authority file of the German National Library) and HLS (Historische Lexikon der Schweiz) entries.

The technology to be used is RDF triple store. This is a NoSQL (non-relational) mechanism for storing and retrieval of data. In a triple store each data entity is composed of subject-predicated-object (triple), like "John knows Mary". An RDF triple store can be viewed as a graph, where an object of one entity is a subject of another, and so on. Graph data representation is particularly pertinent for Digital Humanities where the data is highly interconnected. On practice this kind of data representation guaranties fast path-walking for complex queries enabling knowledge discovery. Furthermore, an RDF triple store has a simple and uniform standard data model, and is governed by a powerful standard query language SPARQL. An RDF triple store also provides a standardized interchange format (e.g. N-triples) for import/export which is important for data transfer/exchange. Thus, RDF triple store is a mature, stable technology convenient for persistent data storage. Moreover, RDF is a standard model for data interchange over the emerging Semantic Web and Linked Open Data cloud. As a future goal, we aim to integrate our RDF data into other international projects (e.g. DBpedia, Europeana) for a higher visibility. Another future direction would be participation in such "meta"-projects as, for example, Bibliographie-Portal (http://www.biographie-portal.eu).