Database of historical places, persons, and lemmas

Natalia Korchagina
Outline

1. Introduction
   1.1 Swiss Law Sources Foundation as a Digital Humanities project
   1.2 Data to be stored
   1.3 Final goal: how it all should be at the end

2. Database technology choice
   2.1 Relational vs. non-relational databases in a DH project

3. Current prototype: RDF triple store
   3.1 Features of triple stores
   3.2 Schema
   3.3 External references
   3.4 Controlled vocabularies

4. Web application for the prototype
   4.1 The WSGI app & SPARQL-endpoint
   4.2 Interface

5. Next steps
1. Introduction

- Digital Humanities: from digitisation of the printed materials to *Linked Data* and *Semantic Web*

- Efficient cataloging and storing of humanities data facilitating research and knowledge exchange within the field

- What is usually stored in a DH project:
  - documents as objects (e.g. books, letters, paintings)
  - organised information on objects (e.g. about certain places, persons)
1. Introduction

- Digital Humanities: from digitisation of the printed materials to *Linked Data* and *Semantic Web*

- Efficient cataloging and storing of humanities data facilitating research and knowledge exchange within the field

- What is usually stored in a DH project:
  - documents as objects (e.g. books, letters, paintings)
  - *organised information on objects* (e.g. about certain places, persons)
1.1 Introduction: Swiss Law Sources Foundation as a Digital Humanities project

- 118 published volumes with critical editions of Swiss historical legal texts
- Texts’ creation time: 10-18th century
- Languages of the collection: German, French, Italian, Romansh, and Latin
- The collection has been scanned and is available online at [http://ssrq-sds-fds.ch/online/](http://ssrq-sds-fds.ch/online/)
- 28 volumes published electronically → full text as FrameMaker/ LaTeX/ Adobe InDesign
- Upcoming volumes are edited in TEI/XML format
1.2 Introduction: data to be stored

- **Back-of-the-book index** of places, persons and lemmas in each volume
- Until recently was composed manually as textual files or Excel tables → constant look-ups for checking whether a given person/place/lemma or its variant/attribute already exists
- **Information** on the same entities is dispersed in various volumes: not practical for search, possible mistakes
1.3 Introduction: final goal

- Database as a tool in the volume edition workflow:
  - An entity is found in the source text (e.g. Hartman ze Chur)
  - Go to the DB’s webpage, look for ‘Hartman’
  - If the right person is found, check the entry of this person
    - If everything is in: tag this person in the text with the corresponding ID
    - Else: add the new info to the DB, then tag the person
    - Else: create a new DB entry on this person

- Database and its GUI to open the foundation’s data on persons/places/lemmas of Switzerland to the large public (historians, linguists and everyone interested)
1.3 Introduction: final goal

- Intuitive GUI with various search options, integrated maps and source texts as PDF for each entry

![Search results for Hartman](image)
1.3 Introduction: final goal

- Intuitive GUI with various search options, integrated maps and source texts as PDF for each entry

Databank of historical places, persons and lemmas

ID: per000377

Names:
1. Hartmann von Werdenberg-Sargans (de) & (it), also known as: Hartmann IV. (1353-1388)
2. Hartmann von Chur (de) & (it), also known as: Hartmann II. (1388-1416)

First mentioned: 1350
Date of death: 16.09.1416
Occupation: Bischof in Chur (1388-1416)
Role: Graf (1350-1416)
Residence: Chur

In source texts also mentioned as:
Hartmann ze Chur

http://dnb.info/grnd/137461318

egenanten kilchen beliben ungesumpt und ungeirt von den egenanten herren von Sangans und iren erben.

Des ze urkündt und stäfer sicherhait, so hab ich, egenanter graf Hainrich, erbetten minen lieben herren und vettern bischoff Hartman ze Chur, das er sin insigel von miner bett wegen hat gehenkt an disen brief, wan ich mins nit by mir han. Wir, Hartman, von gottes1 bischoff, verjechent och, das wir e-unser insigel- durch das egenanten unsers lieben vetters graf Hainrichs bett willen und ze urkündt dir vorges[chribnen] ding, uns, unsern gotzhus und unsern er-
2.1 DB technology choice: relational vs. non-relational

Given and required:

› Highly interconnected data: typical for Digital Humanities

(persons(Person) living in places (Place) being part of other places (Place)/organisations(Org) which have roles (Term) and professions (Term) etc.)

› Non-regular and evolving data structure

› Complicated queries for knowledge discovery, e.g.:

“Find all persons who moved from one residence location to another within a certain period of time”
2.1 DB technology choice: relational vs. non-relational

Relational DB:
- Rigid schema
- Join-intensive query performance deteriorates as the dataset grows bigger
- Struggle with highly connected domains: queries become computationally complex

Non-relational DB:
- Schema is flexible or absent
- Join-intensive query performance remains constant as the dataset grows bigger (at least for some kinds of NoSQL DBs)
- No tabular structure - queries are less expensive
3. Current prototype: RDF triple store

3.1 Features of triple stores:

‣ Simple, uniform data model:
  statements about resources in the form of subject–predicate–object expressions:

```
<table>
<thead>
<tr>
<th>subject</th>
<th>predicate</th>
<th>object</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>has_ID</td>
<td>per0001</td>
</tr>
</tbody>
</table>
```

‣ Powerful, standardised, declarative query language SPARQL

‣ Data **portability** (import/export as N-triples or N-quads)

‣ Data can be represented as a directed labeled **graph** → reflects the specificity of data in Digital Humanities

‣ RDF is a standard model for data interchange over the emerging Semantic Web and Linked Open Data cloud

‣ RDF triples can be integrated into other international projects (e.g. DBpedia, Europeana) → a higher visibility for our data
3. Current prototype

3.2 Schema

- Long process of schema creation

- Different sets of attributes and relationships for different kinds of entities (persons, places, lemmas, organizations)

- 1\textsuperscript{st} step: modelling of objects in XML to visualise the dependencies between different kinds of entities’ features

- 2\textsuperscript{nd} step: unite everything in a schema.

- Though RDF doesn’t require any schema at all, it is mandatory to set one for consequent DB development and consistent use of the DB
XML modelling of a person-entry

<?xml version="1.0" encoding="UTF-8"?>

<person xmlns="per000377"
   std_spelling xmlns="per000377.01">
   <personName xml:lang="de" type="std" gender="1">Hartmann von Werdenberg-Sargans</personName>
   <personName xml:lang="fr" type="std" gender="1">Hartmann de Werdenberg-Sargans</personName>
   <personName xml:lang="it" type="std" gender="1">Hartmann von Werdenberg-Sargans</personName>
   <personSurname xml:id="org000008"/>
   <period from="1355" to="1388" genname="IV."></genname>
   </person>
   </std_spelling>

<std_spelling xmlns="per000377.02">
   <personName xml:lang="de" type="std" gender="1">Hartmann von Chur</personName>
   <personName xml:lang="fr" type="std" gender="1">Hartmann de Coire</personName>
   <personName xml:lang="it" type="std" gender="1">Hartmann von Coira</personName>
   <personSurname xml:id="org000053"/>
   <period from="1388" to="1416" genname="II."></genname>
   </person>
   </std_spelling>

<addname xml:lang="de" type="std" from="""" xmlns="lem/keyXXXXX" />
<addname xml:lang="org" type="orig" from="""" />
<firstMentioned>1350</firstMentioned>
<birth />
<death>16.09.1416</death>
<period from="1388" to="16.09.1416">
<desc xml:lang="de" xml:id="key000015">Bischof</desc>
<location xml:id="loc000021">Chur</location>
</period>
</occupation>

<roleName from="1350" to="16.09.1416" xml:id="key000019">Graf</roleName>
3. Current prototype: Schema
3. Current prototype: Schema

3.3 External references

Entries linked to external, Swiss and international, reference source:

- **HLS**: Historische Lexikon der Schweiz;
- **Idiotikon**: Dictionary of Swiss German;
- **GND**: Integrated authority file of the German National Library;
- **histHub.net**: Swiss metadata hub for historical entries, (planned)
3. Current prototype: Schema
3.4 Controlled vocabularies

It is reasonable to store repeatedly used units:

‣ first names and keywords: professions, types of places and organisations, etc.

in a consistent manner → use of the controlled vocabularies.

‣ Enable a more efficient querying and consistent updating of the database

‣ Collaboration with the Labeling System (i3mainz):
  ‣ adopts the concept of Linked Open Data → complies with the nature of our database
  ‣ vocabularies are RDF graphs available via a REST interface.
4. Web application for the prototype

4.1 The WSGI app & SPARQL endpoint

Initial development settings

Data:

- Index of persons of one particular volume
- Given in format: JSON (converted from manually created Excel-file)
- JSON converted to Turtle for loading into the DB (easy conversion script)
- Contains about 75,000 triples

Application:

- Python script representing a WSGI (Web Server Gateway Interface)
- WSGI interacts with the SPARQL endpoint sending queries wrapped by SPARQLwrapper
- Framework: Flask/Werkzeug
- GUI Templates: results are rendered with Jinja2 templates
4. Web application for the prototype

4.2 Interface

- To date: a prototype at
  
  http://sardona.ssrq-sds-fds.ch:8000

  Achtung! This is still a prototype in development - the interface is error-prone.

- Browsing by persons’ surnames, surname+name, persons’ IDs, organisations’ IDs.

- Some limited filtering options
4. Web application for the prototype

4.2 Interface

<table>
<thead>
<tr>
<th>ID</th>
<th>Vorname</th>
<th>Zuname</th>
<th>Gebur/Ersterw.</th>
<th>Tod/Letsterw.</th>
<th>Beruf/Amt</th>
<th>Wohn/Herkunfts ort</th>
</tr>
</thead>
<tbody>
<tr>
<td>per002207</td>
<td>Ambros</td>
<td></td>
<td>30.01.1679 *</td>
<td>03.08.1738 †</td>
<td>Abt des Klosters Pfäfers (1725-1738)</td>
<td>Pfäfers Kloster</td>
</tr>
<tr>
<td>per002687</td>
<td>Anton</td>
<td></td>
<td>1674</td>
<td></td>
<td>Dekan des Klosters Pfäfers (1674)</td>
<td>Pfäfers Kloster</td>
</tr>
<tr>
<td>per002801</td>
<td>Bartholomäus</td>
<td></td>
<td>1575</td>
<td>1584</td>
<td>Abt des Klosters Pfäfers (1575-1584)</td>
<td>Pfäfers Kloster</td>
</tr>
<tr>
<td>per003113</td>
<td>Basilius</td>
<td>von Rapperswil</td>
<td>19.09.1755 *</td>
<td>08.02.1822 †</td>
<td>Kardinal des Klosters Pfäfers (1796)</td>
<td>Pfäfers Kloster</td>
</tr>
<tr>
<td>per003210</td>
<td>Beda</td>
<td></td>
<td>05.07.1611 *</td>
<td>02.07.1691 †</td>
<td>Abt des Klosters Pfäfers (1637-1644)</td>
<td>Solothurn</td>
</tr>
<tr>
<td>per003240</td>
<td>Benedictus</td>
<td></td>
<td>1673</td>
<td>1722</td>
<td>Dekan des Klosters Pfäfers (1693-1699)</td>
<td>Pfäfers Kloster</td>
</tr>
<tr>
<td>per003112</td>
<td>Benedikt L</td>
<td>der Eide</td>
<td>17.10.1727 *</td>
<td>19.11.1805 †</td>
<td>Administrator (1766-)</td>
<td>Pfäfers Kloster</td>
</tr>
</tbody>
</table>

Public read-only interface
4. Web application for the prototype

4.2 Interface

In the image, there is a screenshot of a web application interface. The interface includes information about a person named Bochler, Benedikt I. (1228-1296). The data includes their standard name, birth and death dates, and various roles they held. The interface is described as a public read-only interface.
4. Web application for the prototype

4.2 Interface

Interface for editors: creation of new entries and update of the existing ones
5. Next steps

- Implementation of controlled vocabularies
- More advanced search options
- **All CRUD operations** (Create-Read-Update-Delete)
- A more evolved user management system & history