Controlled Vocabularies for the Digital Humanities

Dr.-Ing. Michael Piotrowski

Leibniz Institute of European History
<piotrowski@ieg-mainz.de>

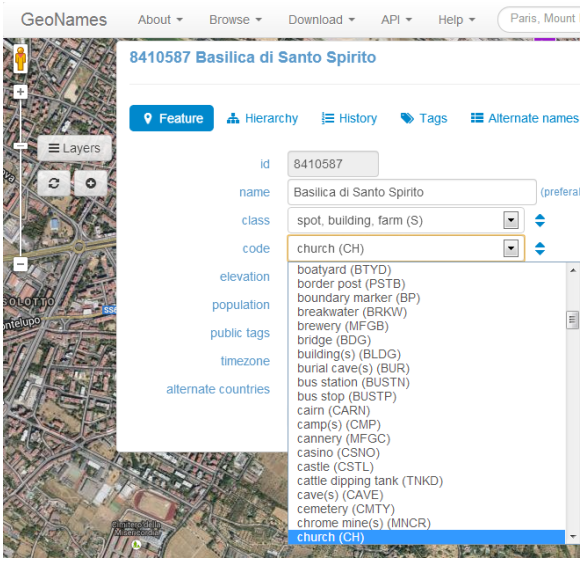
ITUG-Jahrestagung 2013, Mainz · September 16, 2013

Intro

- Digital humanities combine traditional **qualitative** methods with **quantitative**, computer-based methods and tools (e.g., information retrieval, text analytics, data mining, visualization, GIS)
- Quantitative analysis requires **classification**; automatic analysis requires **consistent, unambiguous** classification
- Exchange of data and results requires **shared** classification systems

➜ **Controlled vocabularies**

Example: GeoNames



GeoNames

Collaborative GIS system

Features types

An extensive flat controlled vocabulary, organized into broad categories

Considerations

Accurate? Unambiguous? Discriminative? Consistently applicable? Extensible?

Another example: Library classification

Library of Congress Subject Headings

- History (General)
  - General
  - Military and naval history
  - Political and diplomatic history
  - Ancient History
  - Medieval and modern history, 476–
  - Medieval history
    - Migrations
    - Crusades
    - Latin Kingdom of Jerusalem
    - Later medieval
  - Modern history, 1453–

Dewey Decimal Classification

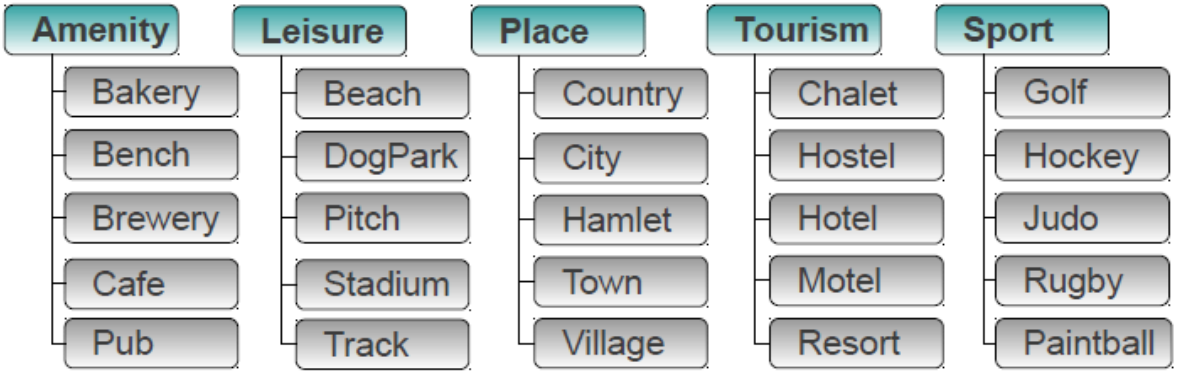
500 Natural sciences and mathematics
510 Mathematics
516 Geometry
516.3 Analytic geometries
516.37 Metric differential geometries
516.375 Finsler Geometry

Considerations

Easy to fit new subjects? What about disciplinary overlaps?

What's a controlled vocabulary?

A controlled vocabulary is a selected list of terms, possibly with definitions, used to categorize things. It supports retrieval and comparison by means of **abstraction** from the detail and **language-independence**.
Can be organized in various ways.



Excerpt from the LinkedGeoData ontology

IEG
Leibniz-Institut für Europäische Geschichte

In the Digital Humanities?

- Vocabularies mostly small-scale and project-specific
- TEI defines names but not values of attributes:

  > *Constructing a list of acceptable attribute values for the @type attribute for each element, on which everyone could agree, is impossible.* (Best Practices for TEI in Libraries, section 3.8.1)

- Popular authority files (e.g., GND, TGN) contain vocabularies, but these are not designed for independent use
  Place type "*region*" in TGN:
  1. Generic geographic region,
  2. Italian administrative entity ("regione"),
  3. and there's a "*generic region*" place type as well…
- Modern vocabularies not suitable for historical research.

IEG
Leibniz-Institut für Europäische Geschichte

What do we need?

- We need controlled vocabularies that go beyond individual projects and enable exchange and collaboration
- Challenges:
  - Realizing descriptive adequacy for the intended application domain
  - Finding the right levels of abstraction and granularity
  - Achieving widespread community agreement

➜ Development of controlled vocabularies must become a community-driven, collaborative endeavor

IEG
Leibniz-Institut für Europäische Geschichte

Our take on it

- Controlled vocabularies are a **focus area** of the IEG in DARIAH
- Our domain: Historical scholarship
- Starting point: A vocabulary of **historical place types** (early modern period, Europe)
- We take the lead, community involvement via experts workshop and DARIAH partners

IEG
Leibniz-Institut für Europäische Geschichte

Design requirements

1. Allow for comparisons of tagged information, among projects and at different level of abstraction (data integration)
2. Interoperability and portability
3. Scalability
4. More accurate retrieval: avoid or manage the ambiguity of natural language (knowledge organization)
5. Automatic reasoning

How?

We need a stricter model to fulfill these requirements. One way could be to use ontologies: simplified but strictly defined formalization of a conceptualization.

What we do at IEG (in the framework of DARIAH)

Ontologies?

Hard to define comprehensively, cumbersome. Furthermore, wouldn't we loose the power of natural-language generic concepts?

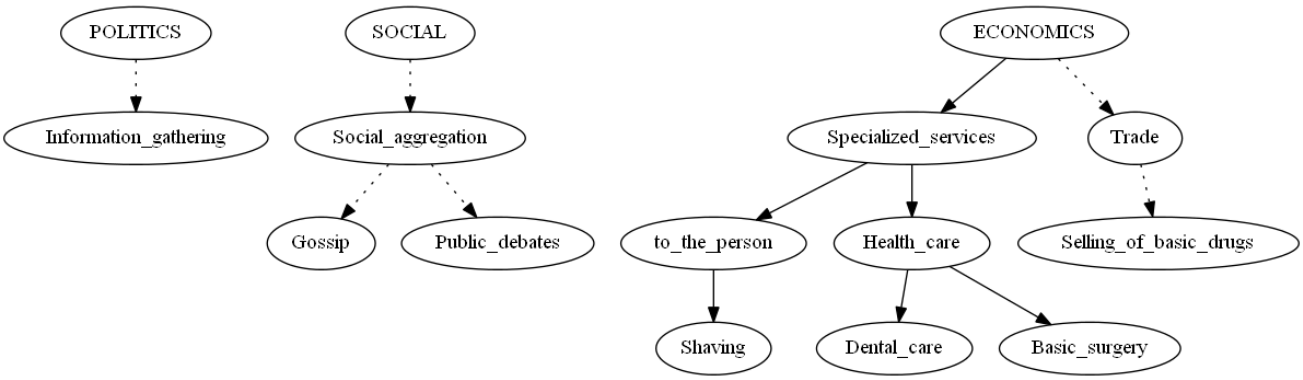
Let's keep them both!

We work on an integrated approach:

- Develop a back-end ontology, which deals with the domain from a high level conceptual perspective, and narrows it down as needed. It must be expressed formally and be throughly documented
- Vocabularies are then built as needed, in natural langage, but associating tags with formally defined concepts

Controlled vocabularies for historical place types

Back-end ontology of functions and actions, natural language tags.
Example: “barber-shop” (Venice, 16th century):



Leaves represent functions associated with the tag. Dotted edges are non-direct relations. Simplified example, not reflecting final version.

IEG
Leibniz-Institut für Europäische Geschichte

Conclusions

- Controlled vocabularies will play a major role in digital humanities, specifically for data integration and knowledge management
- Today’s humanities landscape is littered with project-specific solutions
- TEI has proven that standards can work in the humanities—perhaps it’s time to tackle vocabularies now

IEG
Leibniz-Institut für Europäische Geschichte