



IEG


Leibniz-Institut für
Europäische Geschichte

Mitglied der

Leibniz
Leibniz-Gemeinschaft

The Labeling System

A New Approach to Overcome the Vocabulary Bottleneck

Michael Piotrowski¹ <piotrowski@ieg-mainz.de>  @true_mxp
Giovanni Colavizza² Florian Thiery³ Kai-Christian Bruhn³

¹Leibniz Institute of European History, Mainz, Germany

²DHLab, EPFL, Lausanne, Switzerland

³izmainz, University of Applied Sciences Mainz, Germany

September 16, 2014



Authority files and controlled vocabularies

Natural language is characterized by:

- ▶ homonymy
- ▶ synonymy
- ▶ polysemy

- ▶ Rarely a real problem for humans ...
- ➔ ... but a big problem for semantic information processing, e.g., when annotating and searching



Authority files and controlled vocabularies

Authority files Identification of individuals

Controlled vocabularies Identification of classes



IEG

Leibniz-Institut für
Europäische Geschichte

Authority files and controlled vocabularies

Authority files Identification of individuals

Donnersberg	↪	4012733-3
Donnersberg	↪	4707367-6
Matterhorn	}	↪ 4037992-9
Mont Cervin		
Monte Cervino		
Zugspitze	↪	4068088-5

Controlled vocabularies Identification of classes

Authority files and controlled vocabularies

Authority files Identification of individuals

Donnersberg	↪	4012733-3
Donnersberg	↪	4707367-6
Matterhorn	}	↪ 4037992-9
Mont Cervin		
Monte Cervino		
Zugspitze	↪	4068088-5

Controlled vocabularies Identification of classes

Authority files and controlled vocabularies

Authority files Identification of individuals

Donnersberg	↦	4012733-3
Donnersberg	↦	4707367-6
Matterhorn	}	↦ 4037992-9
Mont Cervin		
Monte Cervino		
Zugspitze	↦	4068088-5

Controlled vocabularies Identification of classes

Donnersberg	}	↦ 4144619-7 (“mountain”)
Matterhorn		
Zugspitze		

Controlled vocabularies

- ▶ Coding relevant properties in uniform and binding form
- ▶ Abstraction from von irrelevant details
- ▶ Abstraction from language-dependent connotations
- Prerequisites for semantic information processing



Names

Last name	First name
Blattmann	Franz Joseph
Kessler	Anton
Lendi	Caspar
Mayer	Balz
Meyer	Gerold
Schorno	Christoff
Stucki	Johannes
Sutter	Beat
Tschudi	Aegidius

Occupations according to sources

Last name	First name	Occupation
Blattmann	Franz Joseph	beck
Kessler	Anton	pfister
Lendi	Caspar	veilpfister
Mayer	Balz	brotbeck
Meyer	Gerold	käsgrempler
Schorno	Christoff	wynziecher
Stucki	Johannes	spend pfister
Sutter	Beat	isenkremer
Tschudi	Aegidius	brodtschätzere

Normalized occupations

Last name	First name	Occupation	Occupation (norm.)
Blattmann	Franz Joseph	beck	Bäcker
Kessler	Anton	pfister	Bäcker
Lendi	Caspar	veilpfister	Backwarenhändler
Mayer	Balz	brotbeck	Bäcker
Meyer	Gerold	käsgrempler	Käsehändler
Schorno	Christoff	wynziecher	Weinhändler
Stucki	Johannes	spend pfister	Bäcker
Sutter	Beat	isenkremer	Eisenwarenhändler
Tschudi	Aegidius	brodtschätzere	Brotbeschauer



Controlled vocabularies

Code	Designation	Definition
001	Bäcker	Artisan making bread
002	Backwarenhändler	...
003	Brotbeschauer	...
004	Eisenwarenhändler	...
005	Käsehändler	...
006	Weinhändler	...

Controlled vocabularies

Code	German designation	English designation
001	Bäcker	Baker
002	Backwarenhändler	Bread dealer
003	Brotbeschauer	Bread inspector
004	Eisenwarenhändler	Hardware dealer
005	Käsehändler	Cheese dealer
006	Weinhändler	Wine dealer

Controlled vocabularies

A	Handwerk	Trades
001	Bäcker	Baker
002	Backwarenhändler	Bread dealer

B	Handel	Commerce
001	Eisenwarenhändler	Hardware dealer
002	Käsehändler	Cheese dealer
003	Weinhändler	Wine dealer

C	Verwaltung	Administration
004	Brotbeschauer	Bread inspector

Shared controlled vocabularies

... enable:

- ▶ reuse
- ▶ semantic data exchange
- ▶ development of generic tools

... require:

- ▶ consensus
- ▶ standardization



Shared controlled vocabularies

... enable:

- ▶ reuse
- ▶ semantic data exchange
- ▶ development of generic tools

... require:

- ▶ consensus
- ▶ standardization



Why are we working on it?

- ▶ Categorization of research data is an integral part of the research process
- ▶ “Right” level of abstraction depends on the research question
- ▶ Shared vocabularies must be accepted by the community
- Development of controlled vocabularies for historical research must be a part of historical research



Why are we working on it?

- ▶ Categorization of research data is an integral part of the research process
- ▶ “Right” level of abstraction depends on the research question
- ▶ Shared vocabularies must be accepted by the community
- Development of controlled vocabularies for historical research must be a part of historical research



Why are we working on it?

- ▶ Categorization of research data is an integral part of the research process
- ▶ “Right” level of abstraction depends on the research question
- ▶ Shared vocabularies must be accepted by the community
- Development of controlled vocabularies for historical research must be a part of historical research



IEG

Leibniz-Institut für
Europäische Geschichte

Why are we working on it?

- ▶ Categorization of research data is an integral part of the research process
- ▶ “Right” level of abstraction depends on the research question
- ▶ Shared vocabularies must be accepted by the community
- ➔ Development of controlled vocabularies for historical research must be a part of historical research



Problems

beck	Baker
brotbeck	Baker
pfister	Baker
spend pfister	Baker
veilpfister	Bread dealer
brodtschätzere	Bread inspector
isenkremer	Hardware dealer
käsgrempler	Cheese dealer
wynziecher	Wine dealer

- ▶ Historical facts are often uncertain: different interpretations may be possible
- ▶ Categories are often closely tied to research questions
- ▶ Shared vocabularies difficult to establish
- ▶ No comparisons accross categories possible

Problems

beck	Baker	?
brotbeck	Baker	?
pfister	Baker	?
spend pfister	Baker	?
veilpfister	Bread dealer	?
brodtschätzere	Bread inspector	?
isenkremer	Hardware dealer	?
käsgrempler	Cheese dealer	?
wynziecher	Wine dealer	?

- ▶ Historical facts are often uncertain: different interpretations may be possible
- ▶ Categories are often closely tied to research questions
- ▶ Shared vocabularies difficult to establish
- ▶ No comparisons accross categories possible

Approach

- ▶ Definition of project-specific vocabularies with references to a reference thesaurus containing “primitive” concepts
- ▶ Bundles of relevant concepts instead of natural language definitions
- Allows for project-specific terminology **and** ensures interoperability (via reference thesaurus)
- Allows for cross-category filtering and comparison

Approach

- ▶ Definition of project-specific vocabularies with references to a reference thesaurus containing “primitive” concepts
- ▶ Bundles of relevant concepts instead of natural language definitions
- Allows for project-specific terminology **and** ensures interoperability (via reference thesaurus)
- Allows for cross-category filtering and comparison

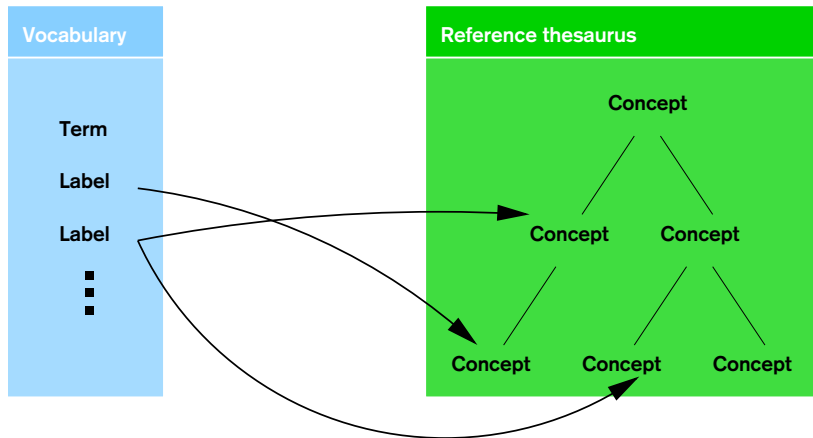
Approach

- ▶ Definition of project-specific vocabularies with references to a reference thesaurus containing “primitive” concepts
- ▶ Bundles of relevant concepts instead of natural language definitions
- ➔ Allows for project-specific terminology **and** ensures interoperability (via reference thesaurus)
- ➔ Allows for cross-category filtering and comparison

Approach

- ▶ Definition of project-specific vocabularies with references to a reference thesaurus containing “primitive” concepts
- ▶ Bundles of relevant concepts instead of natural language definitions
- ➔ Allows for project-specific terminology **and** ensures interoperability (via reference thesaurus)
- ➔ Allows for cross-category filtering and comparison

Approach



Example: concept bundles

Enables “comparing apples to oranges”:

Baker	Bread dealer	Hardware dealer
Bread	Bread	Tools
Production	Distribution	Distribution
⋮	⋮	⋮

Example: concept bundles

Enables “comparing apples to oranges”:

Baker	Bread dealer	Hardware dealer
Bread	Bread	Tools
Production	Distribution	Distribution
⋮	⋮	⋮

Example: concept bundles

Enables “comparing apples to oranges”:

Baker	Bread dealer	Hardware dealer
Bread	Bread	Tools
Production	Distribution	Distribution
⋮	⋮	⋮

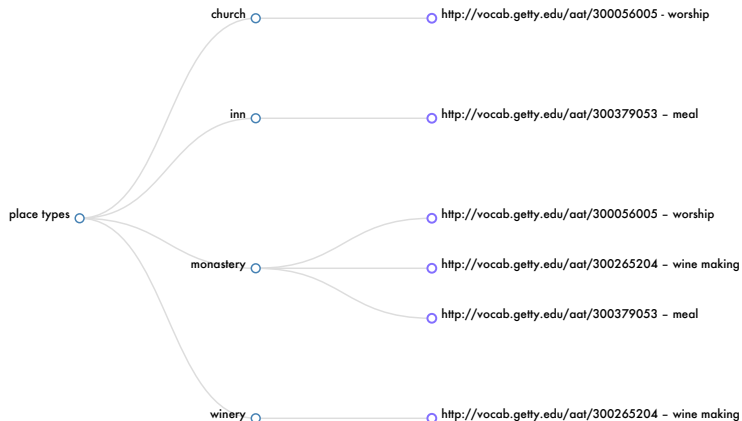
Labeling System

The Labeling System is a tool

- ▶ for creating controlled vocabularies
- ▶ for publishing them (via REST interface and SPARQL endpoint)
- ▶ based on standards (SKOS, RDF, Dublin Core)
- ▶ jointly developed by **IEG** und **i3mainz** within the **DARIAH-DE** framework



Example



Current status

- ▶ **Prototype** at <http://labeling.i3mainz.hs-mainz.de>
- ▶ **Documentation:** http://labeling.i3mainz.hs-mainz.de/share/_ls_doku_v12.pdf

Current work:

- ▶ Migration of the documentation to a wiki
- ▶ Integration of the Labeling Systems into the DARIAH-DE services environment
- ▶ Tutorial for humanities scholars

Conclusion

- ▶ Categorization of research data is an integral part of the research process
- ▶ Controlled vocabularies must be developed within the disciplines
- ➔ Controlled vocabularies (and authority files) are the **intellectual infrastructure** for the digital humanities
- ▶ Concept bundles are a new approach for defining project-specific and at the same time interoperable vocabularies
- ➔ The Labeling System is a tool for creating such vocabularies



IEG


Leibniz-Institut für
Europäische Geschichte

Mitglied der

Leibniz
Leibniz-Gemeinschaft

The Labeling System

A New Approach to Overcome the Vocabulary Bottleneck

Michael Piotrowski¹ <piotrowski@ieg-mainz.de>  @true_mxp
Giovanni Colavizza² Florian Thiery³ Kai-Christian Bruhn³

¹Leibniz Institute of European History, Mainz, Germany

²DHLab, EPFL, Lausanne, Switzerland

³izmainz, University of Applied Sciences Mainz, Germany

September 16, 2014

