Organizing Economic Information

An Overview of Application and Reuse Scenarios of an Economics Knowledge Organization System

_Tobias Rebholz, Andreas Oskar Kempf, Joachim Neubert_
_ZBW – Leibniz Information Centre for Economics – German National Library of Economics_
_Beyond the Numbers, Federal Reserve Bank of St. Louis_
_October 6-8, 2016_
Contents

0. ZBW
1. STW
2. Subject indexing
3. Inter-vocabulary mapping
4. STW for information retrieval support
5. STW web services
ZBW - Leibniz Information Centre for Economics

- World’s largest information and research infrastructure for online/offline economic literature
- More than 4 million volumes and more than 26,000 periodicals and journals
- ECONBIZ - the search engine for economics
- ECONSTOR - publication server for scholarly economic literature
- Application-oriented research in the field of Science 2.0, Web Science, and Knowledge Discovery
- Financing: The ZBW is jointly financed by the federal and state governments
1. STW - Thesaurus for Economics
STW - Thesaurus for Economics

Institutional background
- Developed in cooperation thanks to a project funded by the Ministry for Economy in the 1990s.

Scope
- Covers all economics-related subject areas and the most important related subjects
- Comprises a systematic structure of domain-specific subject categories

Maintenance & Development
- Regularly updated by an editorial team of domain experts from the ZBW
STW - Thesaurus for Economics

Structure
- Polyhierarchial

Languages
- Bilingual: German & English

Types of relations
- Equivalent relations, including synonyms and quasi-synonyms (UF)
- Hierarchical relations, including broader (BT) and narrower terms (NT)
- Associative relations, including related terms (RT)
STW subject categories

Structural characteristics

- The STW subject categories (in total 497) constitute a monohierarchical structure
- Navigation tree
  - Seven main groups » subthesauri
  - Allows thematically browsing in a certain subject field
STW web publication

http://zbw.eu/stw

- Since 2009 published on the web
- Easy reuse
  - Liberal license (ODbL)
  - SKOS format
- URI for each concept
- Readable for humans as well as for machines
- Data is embedded in the HTML web pages via RDFa
2. Subject indexing
Overall indexing situation

- The process of subject indexing is changing more and more into an interplay between various partly interwoven indexing components, in which different indexing methods are applied.

http://www.lkrs.de/images/Startseite/puzzleschieber.jpg
Different indexing scenarios

Incoming documents

Documents without content-descriptive metadata
- Documents in print
  - Intellectual subject indexing (1st level)
- Documents in digital form
  - Generation of standardized c.-d. metadata via machine learning/NLP (2nd level)

Documents with content-descriptive metadata
- Documents with c.-d. metadata from uncontrolled vocabulary
  - Generation of standardized c.-d. metadata via machine learning/NLP (3rd level)
- Documents with c.-d. metadata from controlled vocabularies
  - Intervocabulary mapping via cross-concordances (4th level)
Different indexing scenarios (1st level)

Incoming documents

Documents without content-descriptive metadata
- Documents in print
- Intellectual subject indexing (1st level)

Documents in digital form
- Generation of standardized c.-d. metadata via machine learning/NLP (2nd level)

Documents with content-descriptive metadata
- Documents with c.-d. metadata from uncontrolled vocabulary
- Generation of standardized c.-d. metadata via machine learning/NLP (3rd level)
- Intervocabulary mapping via cross-concordances (4th level)

Documents with c.-d. metadata from controlled vocabularies
Intellectual subject indexing

Intellectual subject indexing by information professionals (1st level):

- Limited to a subset of incoming documents
- Subject metadata is not yet available
- TDM on electronic fulltext is prohibited due to legal restrictions
- Used as the essential basis for further development of the thesaurus
Different indexing scenarios (2nd level)

Incoming documents

- Documents without content-descriptive metadata
  - Documents in print
  - Intellectual subject indexing (1st level)
  - Generation of standardized c.-d. metadata via machine learning/NLP (2nd level)

- Documents with content-descriptive metadata
  - Documents in digital form
  - Documents with c.-d. metadata from uncontrolled vocabulary
  - Generation of standardized c.-d. metadata via machine learning/NLP (3rd level)
  - Documents with c.-d. metadata from controlled vocabularies
  - Intervocabulary mapping via cross-concordances (4th level)
Subject indexing by computers (1)

Subject indexing by computers based on fulltext/abstracts (2\textsuperscript{nd} level):

- Based on text- and data-mining approaches, machine learning algorithms which are able to model human indexing behavior are tested and under active development
- Indexing assistant automatically suggests descriptors
- Quality management, which takes into account the quality of intellectual indexing

Photo: flickr: Hillary - CC BY-SA 2.0
Different indexing scenarios (3\textsuperscript{th} level)

- **Incoming documents**
  - Documents without content-descriptive metadata
  - Documents in print
  - Intellectual subject indexing (1\textsuperscript{st} level)
  - Documents in digital form
  - Generation of standardized c.-d. metadata via machine learning/NLP (2\textsuperscript{nd} level)
  - Documents with content-descriptive metadata
  - Documents with c.-d. metadata from uncontrolled vocabulary
  - Generation of standardized c.-d. metadata via machine learning/NLP (3\textsuperscript{rd} level)
  - Documents with c.-d. metadata from controlled vocabularies
  - Intervocabulary mapping via cross-concordances (4\textsuperscript{th} level)
Subject indexing by computers (2)

Subject indexing by computers based on shorttext (3\textsuperscript{th} level):

- Basic idea: to convert author keywords into STW subject headings based on string match and similarity metrics in combination with/without document title
- Now: Title + keyword indexing with STW
- Allows us to index print publications or publications without the rights for TDM
- 1.7 Million records (title + STW) as training corpus for this method

ZBW - Toepfer, Kempf 2016
Different indexing scenarios (4\textsuperscript{th} level)

Incoming documents

Documents without content-descriptive metadata

- Documents in print
  - Intellectual subject indexing (1\textsuperscript{st} level)

- Documents in digital form
  - Generation of standardized c.-d. metadata via machine learning/NLP (2\textsuperscript{nd} level)

Documents with content-descriptive metadata

- Documents with c.-d. metadata from uncontrolled vocabulary
  - Generation of standardized c.-d. metadata via machine learning/NLP (3\textsuperscript{rd} level)

- Documents with c.-d. metadata from controlled vocabularies
  - Intervocabulary mapping via cross-concordances (4\textsuperscript{th} level)
Cross-concordances

Mappings to other vocabularies (4th level):

- Convert descriptors from other vocabularies into STW descriptors
- Supports truly collaboratively organized subject indexing beyond library boundaries
- Allows an integrated search across various databases indexed with different controlled vocabularies in search portals like EconBiz
- At the beginning mappings were built up exclusively intellectually
- Now: semi-automatic matching procedure
3. Inter-vocabulary mappings
Inter-vocabulary mapping ▶ STW - JEL

Reuse scenario for a JEL – STW mapping effort:

- Economists are usually quite familiar with the JEL classification codes.
- Long-term objective: to animate economists to use STW subject headings in order to provide a more fine-grained content description with a standardized vocabulary.
STW – JEL Mapping (1)

Mapping procedure

- Use of the interactive alignment server AMALGAME (AMsterdam ALignment GenerAtion MEtatool)
- Iterative semi-automatic mapping process
- SKOS vocabulary needed
- Enrichment of STW subject categories and JEL classes
- Exact language dependent string match of STW subject categories and JEL classes.
- Evaluation tool - subsets of alignments can be evaluated manually
STW – JEL Mapping (2)

STW subject categories enriched by
- STW descriptors + synonyms
- Mapped (exactMatch) concepts from other vocabularies – descriptors + synonyms

JEL classes enriched by
- JEL keywords scraped from JEL guide
  https://www.aeaweb.org/jel/guide/jel.php
4. STW for information retrieval support
Index enhancement by search engine

Optimizing the search engine behind EconBiz

- Fast access to the title records is provided by a customized open-source search engine (Solr)
- For the field with STW descriptors, index entries are produced with all synonyms of the descriptors
- Synonym enhancement is active for all “simple” searches
- Main advantage: index-enhancement delivers search results very quickly
- Boosting rules can be applied, which for example rank results higher when a search term appears in the title field of the record or in a thesaurus-enhanced keyword field
5. STW web services
Linked data – STW web services

The ZBW has early on published the STW data additionally in the form of web services

- STW web services execute predefined and pre-optimized SPARQL queries - each service for a particular use case
- Synonym service returns the alternative search terms, optionally including synonyms derived from the mappings to other vocabularies
- Autosuggest service supports data input

http://zbw.eu/beta/econ-ws
Thank you for your attention!

Contact:
Tobias Rebholz
t.rebholz@zbw.eu
Andreas Oskar Kempf
a.kempf@zbw.eu
Joachim Neubert
j.neubert@zbw.eu

http://zbw.eu/stw/

Join the EconBiz Partner Network