KOS evolution in Linked Data

Joachim Neubert
ZBW – Leibniz Information Centre for Economics, Hamburg

SWIB14
Bonn, Germany
03.12.2014
Agenda

- Introduction
- Current versioning approach with STW
- User questions and requirements
- Getting a grip on changes:
  the dataset versioning and skos-history approach
  - Overview
  - Application
  - Selected useful reports
- Outlook: Future work and the skos-history project
STW Thesaurus for Economics

- Created in the 1990s, now maintained and enhanced by ZBW
- More than 6,000 descriptors in English and German
- Since 2009 published as Linked Data in SKOS
- Roughly every year a new version
- Major overhaul in progress – subject area by subject area
Short digression: SKOS as a RDF data format

- Based on *concepts* ("units of thought"), which may bear labels in multiple languages
- All semantic relations (hierarchies, mappings etc.) exist between concepts
- Per language at most one skos:prefLabel (should be unique)
- Additional properties for notations, notes, mappings, etc.
  - Classes for ConceptSchemes and Collections of concepts
- Widely in use today as a common interchange format
How did we handle KOS evolution in the past?
RDF statements about a particular version

<http://zbw.eu/stw>
   a skos:ConceptScheme, void:Dataset ;
   dcterms:issued "2013-10-30"^^xsd:date ;
   owl:versionInfo "8.12" ;
...

Others do this in a similar, yet slightly different way (dcterms:modified, dcterms:hasVersion, ...) – and sometimes, this changes over time
STW versions in URIs

Stable URIs for skos:Concept (and similar for skos:ConceptScheme)

- http://zbw.eu/stw/descriptor/19664-4

303 redirect to versioned URLs (RDFa/rdf/ttl files)


Archived RDFa/rdf/ttl files available


(Currently, search functions and web services always work on the latest version)
Deprecated concepts

No deletion – URI is still defined, shown on a RDFa page like this:

Real estate loan

 Deprecated (used at last in version 8.04), USE Mortgage

<http://zbw.eu/stw/descriptor/12257-3>
  a skos:Concept, zbwext:Descriptor ;
  skos:inScheme <http://zbw.eu/stw> ;
  rdfs:label "Real estate loan"@en, "Realkredit"@de ;
  owl:deprecated true ;
  dcterms:isReplacedBy <http://zbw.eu/stw/descriptor/13775-4> ;
  skos:historyNote "Deprecated (used at last in version 8.04)"@en .
Pragmatic version history solution: Don’t delete anything

STW Thesaurus for Economics

Versions

Prior versions of the STW are provided here for reference (Changes).

Published versions have even version numbers. Odd version numbers are reserved for internal purposes.

- 8.12 (Detailed Changelog - in German)
- 8.10 (Detailed Changelog - in German)
- 8.08 (Detailed Changelog - in German)
- 8.06 (Detailed Changelog - in German)
- 8.04 (first web and linked data version)


Changes are traceable only intellectually (but at all)
Detailed changelog

From legacy maintenance system (simple text file, in German):

Neuangelegte Deskriptoren:

1. Aborigines (Australien) [engl.: Aboriginal Australians] (26584-4)
2. Afghanen [engl.: Afghans] (26068-1)
3. Afghanisch [engl.: Afghan] (26069-6)
4. Afrikaans [engl.: Afrikaans] (26070-0)
5. Afrikaner [engl.: Africans] (26071-5)
6. Afrikanisch [engl.: African] (26072-3)
7. Albaner [engl.: Albanians] (26082-0)
8. Albanisch [engl.: Albanian] (26083-5)
9. Amerikaner [engl.: Americans] (26084-3)
10. Amerikanisch [engl.: American] (26085-1)
11. APEC-Staaten-seitig [engl.: From APEC countries] (26086-6)
12. Araber [engl.: Arabs] (26101-1)
13. Arabisch [engl.: Arab] (26102-6)
15. Aserbaidschanisch [engl.: Azerbaijani] (26129-0)
16. Asiaten [engl.: Asians] (26130-1)
17. Asiatisch [engl.: Asian] (26131-6)
18. Ausländisch [engl.: Foreign] (26132-4)
19. Austauschtheorie (Soziologie) [engl.: Social exchange theory] (25974-3)
How to handle this better?

What users want to know when we publish a new KOS version:

- What’s new?
- What has changed?
Use cases for extended change information

- Human indexers wanting to learn about new and deprecated concepts
- Human indexers (and supporting applications) re-indexing large sets of documents
- People maintaining a derived subset of a KOS
- People maintaining mappings to other vocabularies, and applications supporting them
- Automatic or semi-automatic indexing applications which make use of the KOS and/or its mappings
- Search applications which make use of the KOS and/or its mappings
Getting a grip on changes

(Provided that we have no access to the KOS maintenance system where the changes take place originally, or can’t extend it to report this changes comprehensively.)

Dataset versioning + skos-history

- should basically work on every SKOS vocabulary
5 basic steps to an actionable skos-history

1) Start with a sorted n-triple file per version.
   *(This poses one triple on every single line.)*

2) Create a raw diff between two version files.
   *(This gives you thousands and thousands of differences, even excluding bnodes.)*

3) Split the resulting diff into an insertions and a deletions file.

4) Load the version files, the insertions and deletions files into a triple store as named graphs.

5) Add metadata about the versions and the deltas in a separate „version history graph“.

https://github.com/jneubert/skos-history/blob/master/bin/load_versions.sh
Version History Graph, discoverable via fix URI, e.g.: http://zbw.eu/stw/version
Vocabularies for the plumbing

- dc:/dcterms:
  Dublin Core, as usual the base for everything
- void: http://rdfs.org/ns/void#
  Vocabulary of interlinked datasets
- sd: http://www.w3.org/ns/sparql-service-description#
  SPARQL service description
- delta: http://www.w3.org/2004/delta#
  Differences between RDF graphs
- dsv: http://purl.org/iso25964/DataSet/Versioning#
  Version history records (providing version identifier and date) and a pointer to the current version – outside the actual version data
- sh: http://purl.org/skos-history/
  Scheme and concept version deltas
What’s the benefit?

A database of all versions of a KOS and all deltas between versions – which can be queried in parallel!
SELECT distinct (?concept AS ?addedConcept) (str(?prefLabel) AS ?addedConceptLabel) 
WHERE {
  # parameters
  VALUES ( ?versionHistoryGraph ?language ) {
    ( <http://zbw.eu/stw/version> "en" )
  }
  GRAPH ?versionHistoryGraph {
    # the compared versions default to the current and the previous one
    # get the delta and via that the relevant graphs
    ?delta a sh:SchemeDelta ;
    sh:deltaFrom/dc:identifier ?oldVersion ;
    sh:deltaTo/dc:identifier ?newVersion ;
    sh:deltaFrom/sh:usingNamedGraph/sd:name ?oldVersionGraph ;
    dcterms:hasPart ?insertions .
    ?insertions a sh:SchemeDeltaInsertions ;
    sh:usingNamedGraph/sd:name ?insertionsGraph .
  }
  # for each inserted concept, a newly inserted prefLabel must exist ...
  GRAPH ?insertionsGraph {
  }
  # ... and the concept must not exist in the old version
  FILTER NOT EXISTS {
    GRAPH ?oldVersionGraph {
      ?concept ?p []
    }
  }
  # restrict output to a certain language
  FILTER ( lang(?prefLabel) = ?language )
}
ORDER BY ?prefLabel

Results: Newly inserted concepts

Results of the query:

<table>
<thead>
<tr>
<th>addedConcept</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Accountants</td>
<td></td>
</tr>
<tr>
<td>2 Accounting fraud</td>
<td></td>
</tr>
<tr>
<td>3 Addiction prevention</td>
<td></td>
</tr>
<tr>
<td>4 Alternative currency</td>
<td></td>
</tr>
<tr>
<td>5 Anti-discrimination law</td>
<td></td>
</tr>
<tr>
<td>6 Asset-market approach of the exchange rate</td>
<td></td>
</tr>
</tbody>
</table>
New concepts by subject category

<table>
<thead>
<tr>
<th>Subject Category</th>
<th>Related Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.06 Production management</td>
<td>Cost efficiency</td>
</tr>
<tr>
<td>B.06 Production management</td>
<td>Job performance</td>
</tr>
<tr>
<td>B.06 Production management</td>
<td>Plant asset management</td>
</tr>
<tr>
<td>B.06 Production management</td>
<td>Production process</td>
</tr>
<tr>
<td>B.07 Marketing</td>
<td>Brand architecture</td>
</tr>
<tr>
<td>B.07 Marketing</td>
<td>Brand extension</td>
</tr>
<tr>
<td>B.07 Marketing</td>
<td>Brand loyalty</td>
</tr>
<tr>
<td>B.07 Marketing</td>
<td>Celebrity endorsement</td>
</tr>
<tr>
<td>B.07 Marketing</td>
<td>Consumer attitudes</td>
</tr>
<tr>
<td>B.07 Marketing</td>
<td>Consumer preferences</td>
</tr>
<tr>
<td>B.07 Marketing</td>
<td>Customer acquisition</td>
</tr>
<tr>
<td>B.07 Marketing</td>
<td>Customer integration</td>
</tr>
</tbody>
</table>
## Statistics via aggregation queries: STW

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Added descriptors</th>
<th>Deprecated descriptors</th>
<th>Redirected</th>
<th>Added thsys</th>
<th>Deprecated thsys*</th>
</tr>
</thead>
<tbody>
<tr>
<td>v 8.04</td>
<td>16.02.2009</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v 8.06</td>
<td>22.04.2010</td>
<td>224</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>v 8.08</td>
<td>30.06.2011</td>
<td>131</td>
<td>57</td>
<td>54</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>v 8.10</td>
<td>21.03.2012</td>
<td>105</td>
<td>141</td>
<td>110</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>v 8.12</td>
<td>30.10.2013</td>
<td>260</td>
<td>487</td>
<td>485</td>
<td>12</td>
<td>26</td>
</tr>
<tr>
<td>v 8.14</td>
<td>18.11.2014</td>
<td>227</td>
<td>342</td>
<td>342</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

* Computed column - deprecation and redirects for thsys will be introduced for STW v 8.14 (retrospectively)

https://github.com/jneubert/skos-history/blob/master/bin/create_change_statistics.pl
Statistics via aggregation queries: TheSoz

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Added concepts</th>
<th>Deleted concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>v 0.7</td>
<td>11.01.2011</td>
<td></td>
<td></td>
</tr>
<tr>
<td>v 0.86</td>
<td>08.11.2011</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>v 0.91</td>
<td>30.04.2012</td>
<td>240</td>
<td>4</td>
</tr>
<tr>
<td>v 0.92</td>
<td>19.09.2012</td>
<td>15</td>
<td>3</td>
</tr>
<tr>
<td>v 0.93</td>
<td>25.02.2014</td>
<td>42</td>
<td>4</td>
</tr>
</tbody>
</table>

Thesaurus for the Social Sciences


https://github.com/jneubert/skos-history/blob/master/bin/create_change_statistics.pl
Selected useful reports

- Changed notations
- Splits and merges of concepts
- History of a single concept
## Changed notations (general case)

<table>
<thead>
<tr>
<th>old</th>
<th>new</th>
<th>concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.02.02.01</td>
<td>B.02.03</td>
<td><a href="http://zbw.eu/stw/thsys/71044">http://zbw.eu/stw/thsys/71044</a></td>
</tr>
<tr>
<td>B.06.02</td>
<td>B.06.03</td>
<td><a href="http://zbw.eu/stw/thsys/70310">http://zbw.eu/stw/thsys/70310</a></td>
</tr>
<tr>
<td>B.06.03</td>
<td>B.06.02</td>
<td><a href="http://zbw.eu/stw/thsys/70471">http://zbw.eu/stw/thsys/70471</a></td>
</tr>
<tr>
<td>B.11</td>
<td>B.01.07</td>
<td><a href="http://zbw.eu/stw/thsys/179318">http://zbw.eu/stw/thsys/179318</a></td>
</tr>
<tr>
<td>N.04.04.01</td>
<td>N.04.04.05</td>
<td><a href="http://zbw.eu/stw/thsys/73365">http://zbw.eu/stw/thsys/73365</a></td>
</tr>
<tr>
<td>N.04.04.02</td>
<td>N.04.04.06</td>
<td><a href="http://zbw.eu/stw/thsys/73364">http://zbw.eu/stw/thsys/73364</a></td>
</tr>
<tr>
<td>N.04.04.04</td>
<td>N.04.04.02</td>
<td><a href="http://zbw.eu/stw/thsys/73362">http://zbw.eu/stw/thsys/73362</a></td>
</tr>
<tr>
<td>N.05.08</td>
<td>N.05.07</td>
<td><a href="http://zbw.eu/stw/thsys/73333">http://zbw.eu/stw/thsys/73333</a></td>
</tr>
<tr>
<td>N.05.08.01</td>
<td>N.05.07.01</td>
<td><a href="http://zbw.eu/stw/thsys/73332">http://zbw.eu/stw/thsys/73332</a></td>
</tr>
<tr>
<td>N.05.08.02</td>
<td>N.05.07.02</td>
<td><a href="http://zbw.eu/stw/thsys/73331">http://zbw.eu/stw/thsys/73331</a></td>
</tr>
</tbody>
</table>
## Changed notations (linking STW versioned pages)

<table>
<thead>
<tr>
<th>old</th>
<th>new</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.02.02.01 Capital Budgeting</td>
<td>B.02.03 Capital budgeting</td>
</tr>
<tr>
<td>B.06.02 Manufacturing Systems and</td>
<td>B.06.03 Production organization</td>
</tr>
<tr>
<td>Manufacturing Technology</td>
<td></td>
</tr>
<tr>
<td>B.06.03 Computer-Aided Manufacturing</td>
<td>B.06.02 Factor input</td>
</tr>
<tr>
<td>B.11 Global Management</td>
<td>B.01.07 Global Management</td>
</tr>
<tr>
<td>N.04.04.01 Security Policy</td>
<td>N.04.04.05 Security Policy</td>
</tr>
<tr>
<td>N.04.04.02 International Conflicts</td>
<td>N.04.04.06 International Conflicts</td>
</tr>
<tr>
<td>N.04.04.04 European Integration and EU Policy</td>
<td>N.04.04.02 European integration</td>
</tr>
<tr>
<td>N.05.08 International Law</td>
<td>N.05.07 International law</td>
</tr>
<tr>
<td>N.05.08.01 Public International Law</td>
<td>N.05.07.01 Public International law</td>
</tr>
<tr>
<td>N.05.08.02 Community Law</td>
<td>N.05.07.02 Community law</td>
</tr>
</tbody>
</table>

B.02.02.01 Capital Budgeting

B.02.02.01 Investitionsplanung und -rechnung
(german)

broader

- B.02.02 Corporate Investment Behaviour

Descriptors

- Capital budgeting
- Capital good
- Corporate investment decision
- Discounting
- Disinvestment
- Dynamic capital budgeting
- Foreign direct investment
- Investment
- Investment risk
- Investment theory by the firm
- Mathematics of finance
- Qualitative data analysis
- Real option
- Replacement investment
- Sensitivity analysis
- Useful life

Persistent Identifier (for bookmarking and linking)

- http://zbw.eu/stw/thsys/71044
B.02.03 Capital budgeting

B.02.03 Investitionsrechnung  (german)

broader

- B.02 Corporate finance and capital budgeting

Descriptors

- Business mathematics
- CAPM
- Corporate investment theory
- Decision under uncertainty
- Disinvestment
- Dynamic investment appraisal
- Foreign investment
- Greenfield investment
- Investment
- Investment appraisal techniques
- Investment decision
- Investment risk
- Investment under uncertainty
- Net present value method
- Real options analysis
- Reinvestment
- Sensitivity analysis
- Useful life

Persistent Identifier (for bookmarking and linking)

- http://zbw.eu/stw/thsys/71044
Merges and splits of concepts

... can be recognized by tracing the movement of labels
New concepts, split from old ones

Labels moved to added concepts:

<table>
<thead>
<tr>
<th>oldConcept</th>
<th>movedLabel</th>
<th>newConcept</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commercial professions</td>
<td>Buchhaller</td>
<td>Accountants</td>
</tr>
<tr>
<td>Commercial professions</td>
<td>Accountants</td>
<td>Accountants</td>
</tr>
<tr>
<td>Free-money theory (S. Gesell)</td>
<td>Freigeld</td>
<td>Alternative currency</td>
</tr>
<tr>
<td>Free-money theory (S. Gesell)</td>
<td>Komplementärgeid</td>
<td>Alternative currency</td>
</tr>
<tr>
<td>Free-money theory (S. Gesell)</td>
<td>Regiogeld</td>
<td>Alternative currency</td>
</tr>
<tr>
<td>Free-money theory (S. Gesell)</td>
<td>Regionalgeld</td>
<td>Alternative currency</td>
</tr>
<tr>
<td>Free-money theory (S. Gesell)</td>
<td>Regionalwährung</td>
<td>Alternative currency</td>
</tr>
<tr>
<td>Free-money theory (S. Gesell)</td>
<td>Schwundgeld</td>
<td>Alternative currency</td>
</tr>
<tr>
<td>Labour market discrimination</td>
<td>Affirmative action</td>
<td>Anti-discrimination law</td>
</tr>
</tbody>
</table>

Concept removed and merged into multiple

Minor split-ups of concepts can be revealed by label movements, too:

**Snack food**

*Deprecated (used at last in version 8.12), USE*  
Convenience food

<table>
<thead>
<tr>
<th>deprecatedConcept</th>
<th>movedLabel</th>
<th>newConcept</th>
</tr>
</thead>
<tbody>
<tr>
<td>Snack food</td>
<td>Karloffelchips</td>
<td>Potato</td>
</tr>
<tr>
<td>Snack food</td>
<td>Käsegebäck</td>
<td>Bakery product</td>
</tr>
<tr>
<td>Snack food</td>
<td>Salzgebäck</td>
<td>Bakery product</td>
</tr>
<tr>
<td>Snack food</td>
<td>Schokoriegel</td>
<td>Chocolate</td>
</tr>
</tbody>
</table>

Change history of a concept: “Personnel selection”

```
<http://zbw.eu/stw/descriptor/12571-4/version/8.06/delta/8.08>
  a sh:ConceptDelta;
  dcterms:isPartOf <http://zbw.eu/stw/version/8.06/delta/8.08> ;
  delta:deletion [] ;
  delta:deletion [] ;
  delta:deletion [] ;
  delta:deletion [ skos:altLabel "Job matching"@en ] ;
  delta:deletion [] ;
  delta:deletion [] ;
  delta:deletion [] .

  a sh:ConceptDelta;
  dcterms:isPartOf <http://zbw.eu/stw/version/8.08/delta/8.10> ;
  delta:deletion [] ;
  delta:insertion [ skos:altLabel "Eignungsdiagnostik"@de ] ;
  delta:insertion [ skos:altLabel "Bewerberauswahl"@de ] ;
  delta:insertion [] ;
  delta:insertion [] .

  a sh:ConceptDelta;
  delta:insertion [] .

  a sh:ConceptDelta;
  dcterms:isPartOf <http://zbw.eu/stw/version/8.04/delta/8.06> ;
  delta:deletion [] ;
  delta:insertion [ skos:altLabel "Bewerbungsgespräch"@de ] .
```

http://zbw.eu/stw/descriptor/12571-4
Future work

• For STW:
  • Create a web service for concept history and link a history report to every concept
  • Provide drilldowns for new/deprecated/… concepts from the category level, perhaps visualizations / heat maps

• For skos-history:
  • Apply to differing concept schemes
  • Distill general properties useful for human-readable change reports as well as machine-actionable data
Consider joining the skos-history project …

… particularly if
---
- you are in charge of a KOS and want to publish its change history
- you are using one or several KOS in an application, or intellectually, and want to trace and re-apply upstream changes
- just feel challenged by the task

Code, issues, wiki pages etc.: https://github.com/jneubert/skos-history

Currently, Johan DeSmedt (Tenforce) , Sini Pessala (National Library of Finland) and Agis Papantoniou (Tenforce) are involved in the project and in discussions on which this presentation was based.
Thanks for listening!

Joachim Neubert
ZBW – Leibniz Information Centre for Economics
j.neubert@zbw.eu

http://zbw.eu/stw
https://github.com/jneubert/skos-history
http://zbw.eu/labs