Linked Data als Dauerbaustelle – das Beispiel des STW

Joachim Neubert
ZBW – Leibniz Informationszentrum Wirtschaft
DINI KIM Workshop
Mannheim
11.04.2012

Standard-Thesaurus Wirtschaft

- wurde in den 1990er Jahren unter öffentlicher Förderung von vier wirtschaftswissenschaftlichen Institutionen entwickelt
- wird heute von der ZBW gepflegt und weiterentwickelt
- umfasst über 6.000 Deskriptoren in Deutsch und Englisch
- bildet feinmaschiges Begriffsnetz
  - mehr als 16.000 Ober-/Unterbegriffsbeziehungen
  - fast 11.000 verwandte Begriffe
  - zusätzlicher Zugang über Thesaurussystematik mit über 500 Stellen

Agenda

1) Fortschreiben des STW-Datensets und Versionierung
2) Mapping zwischen Datensets
   a) generelles Verfahren
   b) Erstellen und Fortschreiben von Mappings zu DBpedia
   c) Erstellen und Fortschreiben von Mappings zu SWD
3) Perspektive

STW-Versionen

8.04 – 29.2.2009

- initiale SKOS-Version
- Mapping zu DBpedia

8.06 – 22.4.2010

- breitgestreute Ergänzungen

8.08 – 30.6.2011

- Aufnahme von Mappings Agrovoc, SWD, TheSoz
- Überarbeitung Organisationsforschung, Personalmanagement

8.10 – 21.3.2012

- Überarbeitung Information und Kommunikation

RDF-Statements über STW-Version

```
<http://zbw.eu/stw>
    a skos:ConceptScheme, void:Dataset ;
    dcterms:issued "2012-03-21"^^xsd:date ;
    owl:versionInfo "8.10" ;
...
```

Änderungen gegenüber Vorversion

8.06

- 223 neue Deskriptoren (skos:Concept mit skos:prefLabel de/en)
- 4 gelöschte Deskriptoren
- 618 neue Nichtdeskriptoren (skos:altLabel)

8.08

- 58 neue Deskriptoren
- 57 gelöschte Deskriptoren
- 491 neue Nichtdeskriptoren

8.10

- 105 neue Deskriptoren
- 141 gelöschte Deskriptoren
- 1490 neue Nichtdeskriptoren

Weitere Arten von Änderungen

Insbesondere für das Erstellen/Fortschreiben von Mappings können weiterhin relevant sein (Anzahl aus v8.10):

- Änderung der Vorzugsbenennung (skos:prefLabel) (262)
  „Software-Agent“ -> „Agentenbasierte Modellierung“
- Termänderungen Nichtdeskriptoren (skos:altLabel) (429)
  „Bankkonkurs“ -> „Bankenkonkurs“
- Umgehängte Nichtdeskriptoren (540)
  „Agglomerationsprozess“ von „Regionale Konzentration“ (gelöscht) zu „Zentralisierung“
- Relationenänderung (UB, OB, VB) an Deskriptoren (547)

STW-Versionierungskonzept

stabile URIs für skos:ConceptScheme und skos:Concept

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

303-Redirect auf versionierte URL (Datei, ggf. mit Sprachpräferenz)

- http://zbw.eu/stw/versions/latest/about
- http://zbw.eu/stw/versions/latest/descriptor/19664-4/about

Kompletter Bestand historischer RDFa/rdf/ttl-Dateien wird vorgehalten (mit Hinweis auf obsolete Version)

- http://zbw.eu/stw/versions/8.06/about
- http://zbw.eu/stw/versions/8.06/descriptor/19664-4/about

Suchfunktion und Webservices arbeiten immer auf aktueller Version

Problem aus Semantic Web Perspektive

RDF-Aussagen über einzelne Versionen von Concepts und ConceptScheme (und damit über die Änderungen/Mappings zwischen Versionen) sind nicht möglich

auch in W3C (2009) offene Frage

Versionierte Namespace-URI? – warnende Beispiele:

`http://xmlns.com/foaf/0.1/` für FOAF 0.98

`http://www.w3.org/2004/02/skos/core#` für SKOS Rec 18.8.2009

Proxies? (Ansätze bei Dublin Core):

```
<http://dublincore.org/documents/dcmi-terms/>
    dcterms:modified `2010-10-11`
<http://purl.org/dc/terms/creator> dcterms:hasVersion
    <http://dublincore.org/usage/terms/history/#creatorT-002>
```

Pragmatische Lösung – Versionsübersicht

Standard-Thesaurus Wirtschaft

Versionen

Hier finden Sie frühere Versionen des STW (Änderungen).

Die publizierten Versionen tragen gerade Versionsnummern, die ungeraden Versionsnummern sind für interne Zwecke reserviert.

- 8.10 (Detaillierte Änderunghistorie)
- 8.08 (Detaillierte Änderunghistorie)
- 8.06 (Detaillierte Änderunghistorie)
- 8.04 (erste Web-/Linked-Data-Version)

Bitte benutzen Sie zum Verlinken auf die Thesarusbegriffe die versions- und sprach-unabhängigen URIs (z. B. `http://zbw.eu/stw/descriptor/19664-4` anstelle von `http://zbw.eu/stw/versions/latest/descriptor/19664-4/about.de.html`).

nur intellektuell – aber immerhin überhaupt – nachvollziehbar

Detaillierte Änderungshistorie

übernommen aus STW-Pflegesystem (einfache Textdatei):

```
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)
   14.  Armenier        [engl.: Armenians]      (26103-4)
   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)
```

Gelöschte Deskriptoren/Konzepte

URI weiter definiert – RDFa-Seite:

Realkredit

Stillgelegt (zuletzt verwendet in Version 8.04), BENUTZE
Hypothek

```
<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,
    "Stillgelegt (zuletzt verwendet in Version 8.04)"@de .
```

Mapping zwischen Datensets

Ontology alignment life cycle (Euzenat/Le Duc 2012)

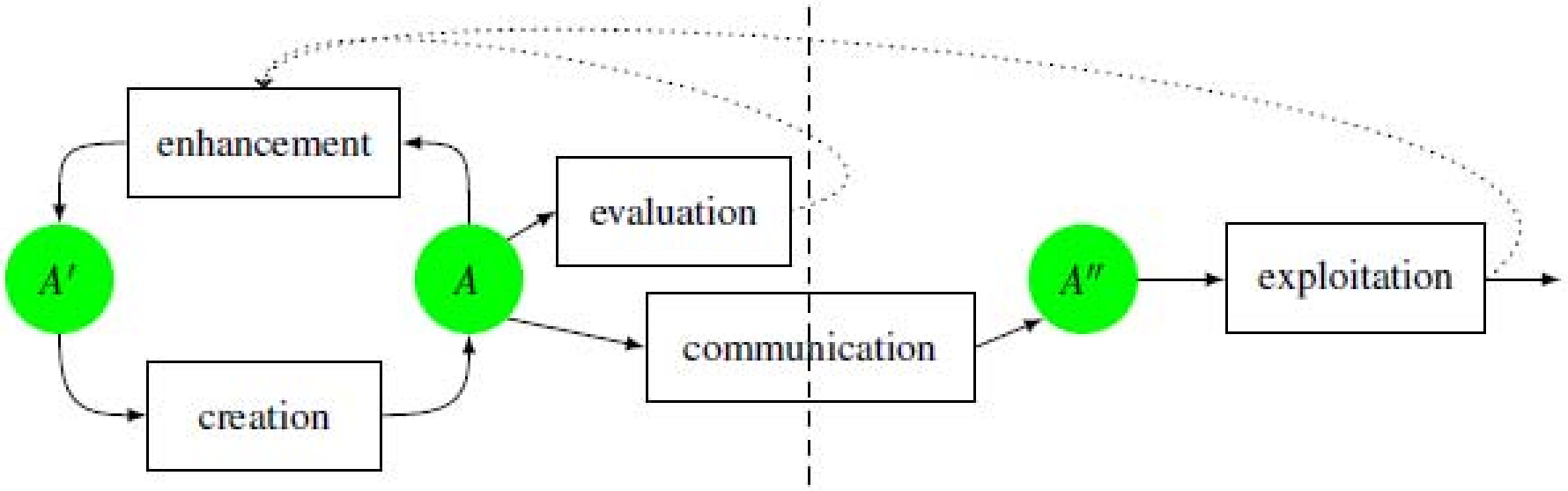


Fig. 2 The ontology alignment life cycle, adapted from (Euzenat et al. 2008).

generelles Verfahren für alle Mappings

Mappingzielspezifisches Preprocessing

- Mapping selbst erstellen oder
- aus 3rd-Party Quelldateien filtern (dabei Auflösung von skos-xl Konstruktionen und sonstigen Besonderheiten der Quellen)

Mapping-Dateien mit reduzierten Statements

- skos:exactMatch/closeMatch/narrowMatch/broadMatch
- skos:prefLabel
- skos:inScheme

Anreicherung mit Statistik und Metadaten (selbstbeschreibend)

Einheitliche Generierung von RDFa- und Download-Files

Standard-Thesaurus Wirtschaft (STW): Mapping zu Schlagwortnormdatei (SWD)

Über das Mapping

Beschreibung:	Aufgebaut durch DNB, USB Köln und ZBW und laufend fortgeschrieben von Expertinnen aus ZBW und DNB
Urheber:	Deutsche Nationalbibliothek (DNB) und ZBW - Leibniz-Informationszentrum Wirtschaft
Lizenz:	http://creativecommons.org/publicdomain/zero/1.0/
Rechte:	Die CC0-Lizenzierung des Mappings soll eine möglichst breite und einfache Wiederverwendung ohne rechtliche Einschränkungen fördern. Über einen Hinweis auf die oben angegebene Urheberschaft und eine freie Verfügbarkeit von Projekten, die dieses Mapping verwenden, würden wir uns jedoch freuen. http://opendatacommons.org/norms/odc-by-sa/
Relationen:	4608 skos:exactMatch 5528 skos:narrowMatch 63 skos:broadMatch
Herausgeber:	ZBW - Leibniz-Informationszentrum Wirtschaft

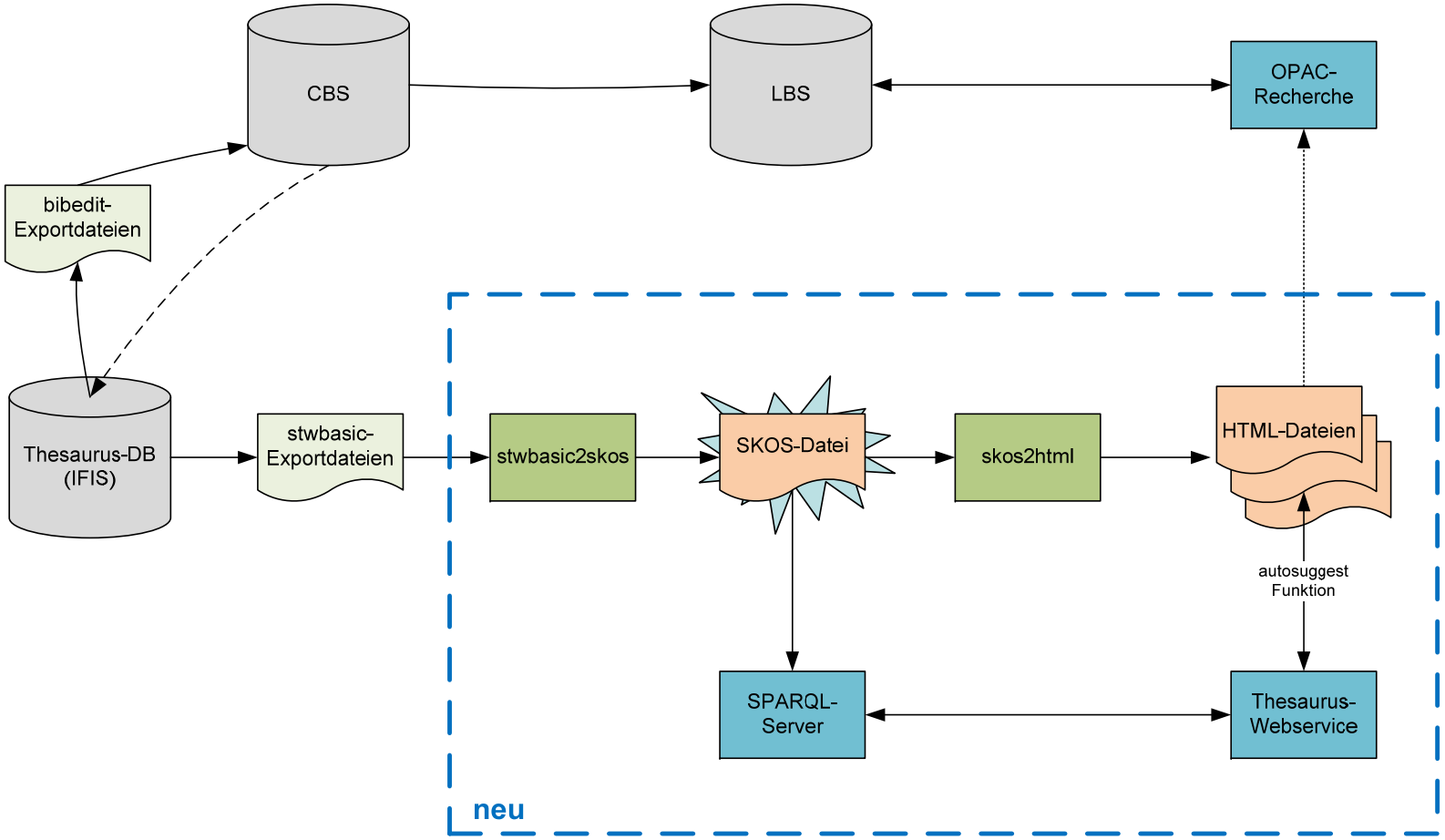
Über Schlagwortnormdatei (SWD)

Beschreibung:	Die Schlagwortnormdatei ist ein kontrolliertes Schlagwortsystem, das vor allem zur Sacherschließung in deutschen Bibliotheken eingesetzt wird.
Urheber:	Deutsche Nationalbibliothek in Kooperation mit verschiedenen Bibliotheken und Bibliotheksverbünden
Homepage:	http://www.d-nb.de/standardisierung/normdateien/swd.htm
Version:	2011-07

RDF-Statements über STW – DBpedia Mapping

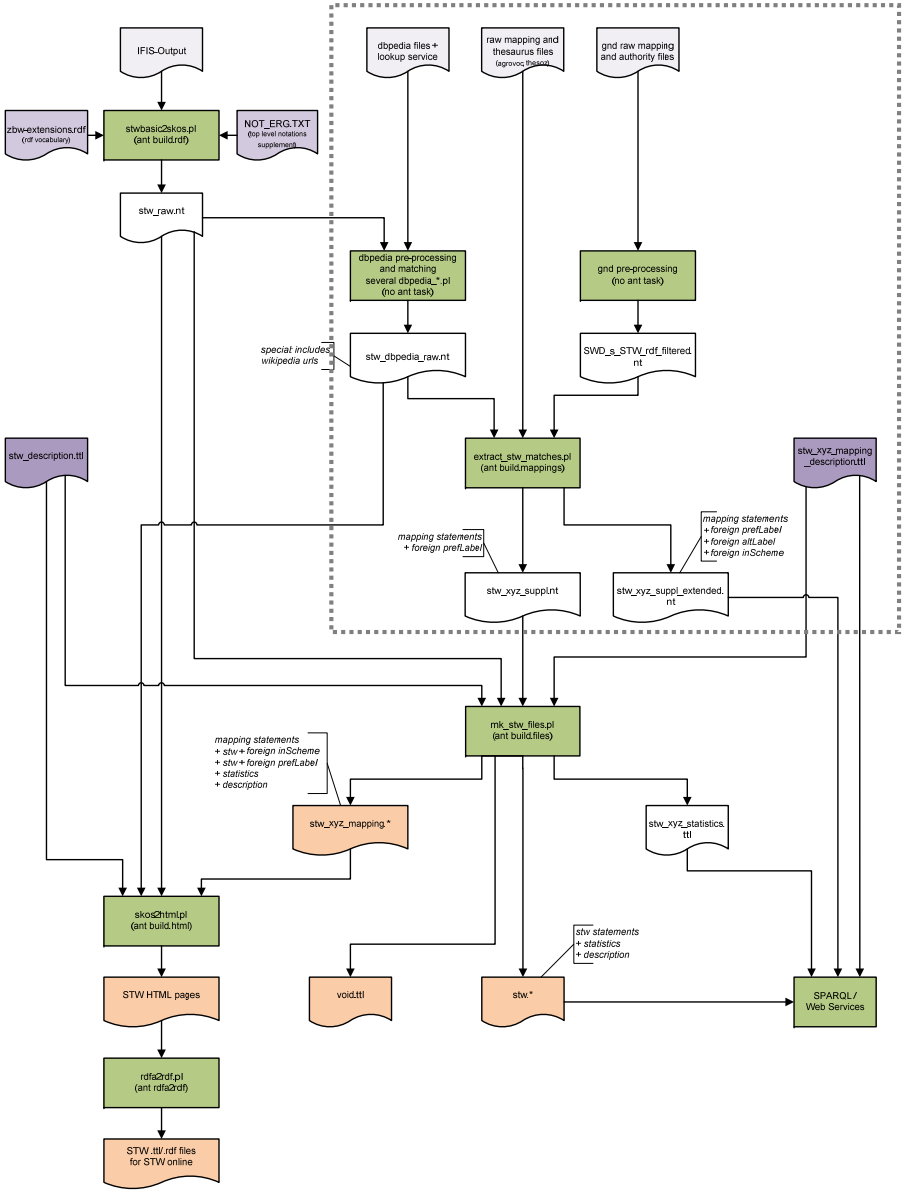
```
<http://zbw.eu/stw/mapping/dbpedia>
    a void:Linkset ;
    void:target <http://zbw.eu/stw>,
        <http://zbw.eu/stw/mapping/dbpedia/target> ;
    void:propertyPartition [
        void:property skos:exactMatch ;
        void:triples "1004"
    ], [
        void:property skos:closeMatch ;
        void:triples "1915"
    ] .

<http://zbw.eu/stw/mapping/dbpedia/target>
    a void:Dataset ;
    foaf:focus <http://dbpedia.org/> ;
    foaf:homepage <http://dbpedia.org/about> ;
    owl:versionInfo "3.6" .
```

Produktion STW allein (2009)



Produktion aktuelle STW-Version (mit Mappings)

stwbasic2skos



skos2html

für weitere Verarbeitung müssen alle Mappings abgeschlossen sein

Erstellen von STW-Mappings

Zwei diametral entgegengesetzte Ansätze:

- DBpedia:
  Erstellung komplett automatisch
- SWD:
  Erstellung/Pflege komplett intellektuell (Kooperation DNB/ZBW)

Von Dritten übernommen Mappings (TheSoz/Gesis, Agrovoc/FAO) werden hier nicht im Detail betrachtet.

Verfahren STW – DBpedia Mapping

Lexikalisches Matching (lowercased labels)

- Nutzung von skos:altLabels / DBpedia Redirects
- Nutzung von deutschen und englischen Labels / verknüpften DBpedia Einträgen

automatische Evaluierung der Matches

- de + en skos:prefLabel matchen beide
  => skos:exactMatch
- nur ein skos:prefLabel (+ ggf. skos:altLabels)
  => skos:closeMatch
- Matches zu unterschiedlichen DBpedia Concepts
  => kein Match

Probleme STW – DBpedia Mapping

- keine intellektuelle Evaluierung der Ergebnisse, keine Korrektur, kein Feedback – Datenset wird mit jedem Mappinglauf komplett ersetzt
- DBpedia = englische Wikipedia (nicht verknüpfte deutsche Wikipedia-Einträge fallen raus, keine deutschen Redirects)
- keine Nutzung des DBpedia Kontexts (Bereich Wirtschaft?)
- stellenweise Verzerrung durch „Up-posting“ im STW (Ananas, Avocado etc. Synonyme zu Tropische Früchte)
- fehlende Konzepte in DBpedia (z.B. Agrarpreise, Jugendarbeitslosigkeit, etc.)

Probleme (Forts.) und Lösungansätze

technische Probleme beim DBpedia-Mapping:

- Redirects und Disambiguation Pages unterschiedlich aufgesetzt
- Strukturänderungen DBpedia-Files mit DBpedia-Versionen
- hoher Aufwand an Hauptspeicher und Rechenzeit

Lösungsperspektive:

- Wikidata-Projekt
- internes Projekt zur Evaluierung und Korrektur
- Tooleinsatz, Versionierung von Mappings, Workflowunterstützung

Verfahren STW – SWD Mapping

Crosskonkordanz bereits lange vor Linked-Data-Publikation erstellt

Pflege rein intellektuell:

- Änderungslisten
- Anlegen/Ändern/Löschen der Relationen mit händischer Eingabe von IDs in Win-IBW (DNB-System)
- bisher separater Bestand, künftig in GND

Probleme STW – SWD Mapping

- sehr arbeitsintensiv
- derzeit keinerlei Toolunterstützung
- zeitliche Lücke zwischen Finalisierung der STW-Version und Abarbeitung der Änderungen der Crosskonkordanz
  - Ausgabe aus DNB-System eigener, aufwendiger Arbeitsschritt
- daraus folgend zwei unschöne Alternativen
  - zwei Generierungsläufe für die RDFa-Seiten *einer* STW-Version (mit alten und mit aktualisierten Mappings), oder
  - Warten auf die nächste STW-Version

Perspektive: Toolgestützter Mapping-Workflow

für alle neuen Concepts (eines Source-Vokabulars):

1. automatische Generierung von Vorschlägen
   ggf. Reduzierung auf einen Ausschnitt (obere/untere Thresholds)
2. intellektuelle Bewertung der Vorschläge
3. Speicherung der Entscheidung (Art des Matches oder noMatch)

optional Re-Evaluierung

skos:closeMatch/broadMatch/narrowMatch, noMatch

neue Version: rinse and repeat ...

Toolgestützer Mapping-Workflow (weitere Anf.)

Für den ersten Schritt (automatische Generierung von Vorschlägen):

- Unterstützung von „vielen unterschiedlichen Dimensionen von Ähnlichkeit“ (Ossenbruggen et al. 2011)
- Pluggable Matching Algorithms
- flexible Chains/Rulesets
- Unterstützung bei der Evaluierung unterschiedlicher Varianten

Idealerweise darüber hinaus:

Erzeugung von Vorschlägen/Kommentaren/Kritik durch Endnutzer (als zusätzlicher Input für Mapping-Workflow)

Tools

„Most alignment tools are not designed for the large but shallow vocabularies ... Furthermore, tools provide little support to analyse large sets of correspondences, making it difficult to assess the quality of the generated results.“ (Ossenbruggen et al. 2011)

- SILK (http://www4.wiwiss.fu-berlin.de/bizer/silk/) *– für Massendaten*
- Amalgame (http://semanticweb.cs.vu.nl/amalgame/)

  – am vielverspechendsten, Video unter http://www.youtube.com/watch?v=_yQtUq5sC6M

andere Ansätze s.u.
http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/EquivalenceMining

Literaturhinweise

Euzenat, J., & Le Duc, C. (2012). Methodological guidelines for matching ontologies. *Ontology Engineering in a Networked World*, 257–278. Retrieved from ftp://ftp.inrialpes.fr/pub/exmo/publications/euzenat2012a.pdf

Ossenbruggen, J. van, Hildebrand, M., & Boer, V. de. (2011). Interactive vocabulary alignment. Presented at the Theory and Practice of Digital Libaries, Berlin. Retrieved from http://semanticweb.cs.vu.nl/lod/tpdl2011/paper.pdf

Shvaiko, P., & Euzenat, J. (2012). Ontology Matching: State of the Art and Future Challenges. *IEEE Transactions on Knowledge and Data Engineering*. doi:10.1109/TKDE.2011.253. Retrieved from http://disi.unitn.it/~p2p/RelatedWork/Matching/SurveyOMtkde_SE.pdf

W3C. (2009). White paper Vocabulary Management - Semantic Web Standards. *Semantic Web Standards Wiki*. Retrieved April 11, 2012, from http://www.w3.org/2001/sw/wiki/White_paper_Vocabulary_Management

Vielen Dank

Joachim Neubert
ZBW – Leibniz Informationszentrum Wirtschaft
j.neubert@zbw.eu

http://zbw.eu/stw
http://zbw.eu/labs