

Evaluating Approaches to automatically match Thesauri from different Domains for Linked Open Data

Ahsan Morshed, Benjamin Zopilko, Gudrun
Johannsen, Philipp Mayr, Johannes Keizer

10th NKOS Workshop, 28.-29.09.2011, Berlin

Motivation

Matchings between thesauri provide a bridge between datasets from different domains

- Guiding users in finding relevant concepts in an interdisciplinary manner
- Exposing overlaps / differences between disciplines
- Providing linkages for the Linked Data cloud
- Testing the effectiveness of simple matching approaches

Use Cases

Using Thesauri Matchings for

- Seamless queries for interdisciplinary information
 - e.g. a researcher interested in agricultural science information evaluated with methods of the social sciences
- Pretesting collections to be included in a web portal
- Getting a thesaurus (and annotated data) into the Linked Data cloud

Matching Thesauri in the Semantic Web

- Thesauri in the Linked Data cloud
 - Most commonly available in SKOS format
 - But: representation of explicit relationships differs widely

- Interoperability among datasets in the Semantic Web
 - Ontology Alignment tools (e.g. FALCON-AO, ASMOV)
 - Mostly require conversion to OWL
 - Link Discovery tools (e.g. Silk, SERIMI)

Case Study

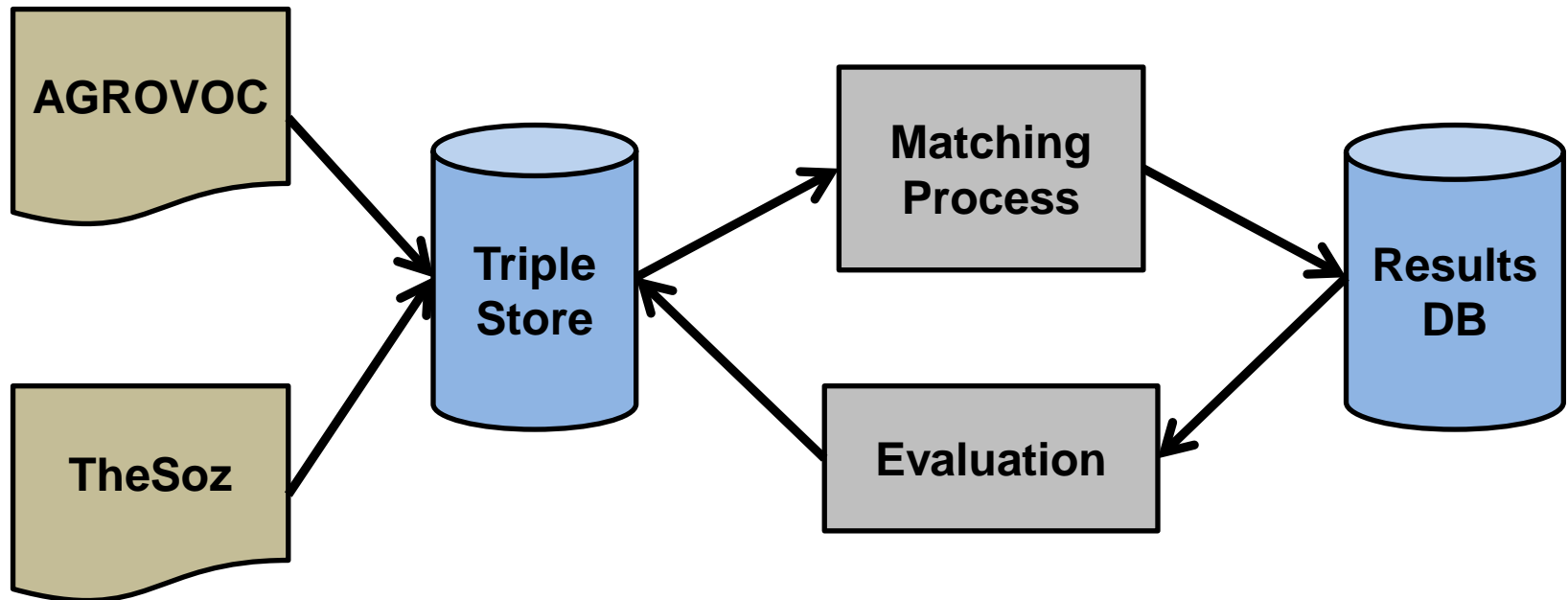
- Task: Evaluating different approaches for matching thesauri automatically
- Involved thesauri should have only few conceptual overlap
- Focus on skos:exactMatch
- Evaluating three groups of approaches
 - Syntactic self-developed algorithm (“Initial Approach”)
 - Link Discovery Tools
 - Ontology Alignment Tools

Involved Thesauri

- AGROVOC from FAO
 - A multilingual agricultural vocabulary
 - Close to 40000 concepts
 - Covers agriculture, forestry, fisheries and related themes (food security, land use, environment, etc.)

- TheSoz from GESIS
 - A multilingual social science thesaurus
 - Ca. 11600 keywords, ca. 7750 descriptors
 - Covers social sciences and related disciplines

Workflow



Algorithm (Initial Approach)

- Storing two SKOS thesauri into a triple store (Sesame)
- Identifying Matches
 - Only preferred labels are considered
 - Levenshtein distance serves as similarity measure
 - Threshold 0.21 is chosen for finding the matches
 - Results are categorized in `skos:exactMatch` and `skos:closeMatch` in order to produce trustful links
- Candidate matches are manually evaluated by a domain expert in a relational database
- Storing correct candidate matches into the triple store

Evaluation Criteria



- Consider non-preferred terms (alternative labels in SKOS terminology) associated with the candidate match term in order to clarify the meaning.
- Consider other languages of the matching terms
- Consider the concept hierarchy, i.e. mainly parent concepts.
- Consider definitions or scope notes of mapped concepts to verify the correctness of exact matches

Results

Matching Approach	# Candidate Exact Matches	# Correct Exact Matches	# Incorrect Exact Matches	Precision
Levenshtein (Initial Approach)	1613	840	773	0.52
Levenshtein (Silk)	288	288	0	1
Normalized Levenshtein (Silk)	660	372	288	0.56

Results AGROVOC

Age groups at AGROVOC Thesaurus
http://aims.fao.org/aos/agrovoc/c_28628

Property	Value
exactMatch	<ul style="list-style-type: none"> <http://aims.fao.org/aos/agrovoc/c_28630> <http://dewey.info/class/305.2/> <http://lod.gesis.org/thesoz/concept/10035257> <http://zbw.eu/stw/descriptor/19760-1>
date of creation	<ul style="list-style-type: none"> 1989-10-05
date of last update	<ul style="list-style-type: none"> 2009-01-29
isInfluencedBy narrower	<ul style="list-style-type: none"> <http://aims.fao.org/aos/agrovoc/c_186> <http://aims.fao.org/aos/agrovoc/c_139> <http://aims.fao.org/aos/agrovoc/c_1547> <http://aims.fao.org/aos/agrovoc/c_28629> <http://aims.fao.org/aos/agrovoc/c_3851> <http://aims.fao.org/aos/agrovoc/c_8491>
prefLabel	<ul style="list-style-type: none"> ALTERSGRUPPE (de) Age groups (en) Groupe d'âge (fr) Grupa wiekowa (pl) Grupo etário (pt) Grupos de edad (es) Gruppi di età (it) korcsoport (hu) vekové skupiny (sk) věkové skupiny (cs) возрастные группы (ru) فئات عمرية (ar) گروه‌های سنی (fa) आयु वर्ग (hi) നേമനായ (th) အသက်အုပ်စုများ (lo) 年齢集団 (ja) 年龄群体 (zh)

Results TheSoz

at GESIS Linked Data Prototype

<http://lod.gesis.org/thesoz/concept/10035257>



Property	Value
is skos:broader of	<ul style="list-style-type: none"> <http://lod.gesis.org/thesoz/concept/10034597> <http://lod.gesis.org/thesoz/concept/10034619> <http://lod.gesis.org/thesoz/concept/10035258> <http://lod.gesis.org/thesoz/concept/10035321> <http://lod.gesis.org/thesoz/concept/10035322> <http://lod.gesis.org/thesoz/concept/10035323> <http://lod.gesis.org/thesoz/concept/10035324> <http://lod.gesis.org/thesoz/concept/10035325>
skos:exactMatch	<ul style="list-style-type: none"> <http://aims.fao.org/aos/agrovoc/c_28628>
skos:narrower	<ul style="list-style-type: none"> <http://lod.gesis.org/thesoz/concept/10034597> <http://lod.gesis.org/thesoz/concept/10034619> <http://lod.gesis.org/thesoz/concept/10035258> <http://lod.gesis.org/thesoz/concept/10035321> <http://lod.gesis.org/thesoz/concept/10035322> <http://lod.gesis.org/thesoz/concept/10035323> <http://lod.gesis.org/thesoz/concept/10035324> <http://lod.gesis.org/thesoz/concept/10035325>
skosxl:prefLabel	<ul style="list-style-type: none"> <http://lod.gesis.org/thesoz/term/10035257>
skos:relatedMatch	<ul style="list-style-type: none"> <http://zbw.eu/stw/descriptor/15905-0>
rdf:type	<ul style="list-style-type: none"> ext:Descriptor

This page shows information obtained from the SPARQL endpoint at <http://lod.gesis.org/thesoz/sparql>.

[As N3](#) | [As RDF/XML](#)

The GESIS Linked Data Prototype uses the [Pubby Linked Data Frontend](#). More information on Linked Open Data at GESIS can be found [here](#).

Outlook

- Expanding the evaluation on additional tools and measures
 - Ontology Alignment tools
 - Link Discovery tools
- Considering term variants in the matching algorithm
- Including more relationships (e.g. broader and narrower)