Ahsan Morshed, Benjamin Zapilko, Gudrun Johannsen, Philipp Mayr, and Johannes Keizer (2011).
*Evaluating approaches to automatically match thesauri from different domains for Linked Open Dat*a.
10th European Networked Knowledge Organisation Systems (NKOS) Workshop, Berlin (Germany)

Evaluating approaches to automatically match thesauri from different domains for Linked Open Data

Ahsan Morshed[1], Benjamin Zapilko[2], Gudrun Johannsen[1], Philipp Mayr[2], and Johannes Keizer[1]

[1] The Food and Agricultural Organization of UN (FAO), Rome, Italy
[2] GESIS – Leibniz Institute for the Social Sciences, Bonn, Germany

Extended Abstract

With the use of SKOS the heterogeneous environment of various vocabularies worldwide can be technically harmonized prospectively and especially the content of traditional databases can be made accessible and connectable for applications of the Semantic Web, i.e. as Linked Open Data. Vocabularies in SKOS format and respectively crosswalks between them can play a relevant role in this context, because they can serve as a bridging hub for the inter-linking of different published and indexed data sets. However, huge effort in developing and evaluating automatic alignment techniques have focused mostly on ontologies in recent years (see activities from the OAEI and van Hage 2008), with the demand that vocabularies in SKOS format would often have to be converted into OWL format.

Our case study presents how thesauri from different domains can be matched automatically and which matching approaches are most promising for this difficult task. Therefore we reprise approaches made in Lauser et al. 2008. The Thesaurus for the Social Sciences (TheSoz) and the AGROVOC thesaurus are established KOS in their domains and by their scope, but they seem to have very few conceptual overlap. Both thesauri are available in SKOS format and are freely available on the web. However, in order to detect possible linkages between both thesauri and to expose them into the LOD cloud, the intention of this paper is to check, if there are any good approaches to find conceptual overlap in thesauri from remote domains (semi-)automatically.

Therefore different approaches for aligning ontologies and linking data sources on the web are performed on both SKOS thesauri. The automatic generated matches, which should preferably be statements with properties skos:exactMatch, skos:closeMatch or owl:sameAs, are then evaluated by domain experts.
The initial matching approach is based on the syntactic algorithm which consists of Levenshtein distance and Jaro Measure. We can adapt it by following the steps (see also Fig. 1):

1. The selected thesauri are downloaded as SKOS resources from their respective websites.
2. A single triple store is created, with all RDF/SKOS triples coming from the thesauri.
3. Each pair of thesauri (AGROVOC-X) is considered at a time, e.g., AGROVOC and TheSoz, and so on).
4. To all of the possible pairs of concepts formed (the first concept coming from AGROVOC, the second one, from the other thesaurus), the following steps are taken:
   - **a.** the preferred label only is considered;
   - **b.** the above similarity measure is applied;
   - **c.** the average of the similarity measure is computed;
   - **d.** a threshold is applied for tuning the measure for finding the matches
   - **e.** mostly skos:exactMatch, and skos:closeMatch are considered.

5. All resulting candidate matches are loaded into a relational database, which are then manually evaluated by a domain expert.
6. Candidate matches that are confirmed by the domain expert are then loaded in the sesame triple store.

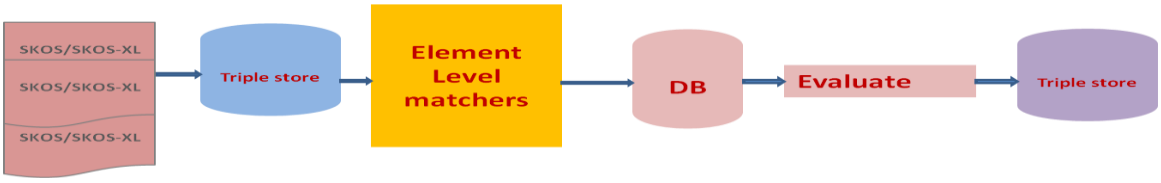

FIG.1. Matching process workflow

In order to compare the results of the initial algorithm, we also evaluate some existing matching tools (i.e. FALCON-AO, COMA++, Silk). Only few of the tools can directly be executed on data in SKOS format. Because most approaches have been originally developed for aligning ontologies, especially those participating in the OAEI, the approaches have either to be adjusted or at least the thesauri have to be converted into OWL. Efforts for these adjustments have to be taken into account.

The matching results of all approaches syntactic and semantic are intellectually assessed concerning their mapping quality. Pros and cons of the approaches are displayed and discussed (compare Lauser et al., 2008). Overlaps in the matching results of the different approaches are identified and interpreted. Furthermore, these mapping links extend the communication among the thesauri and bootstrapping the linked data vision.

As part of the evaluation of the different matching approaches we try to give a recommendation, which approach is suitable for our outlined task and which approaches hold difficulties when being applied on SKOS thesauri.

References

1. Jaro, M. A. (1989). "Advances in record linkage methodology as applied to the 1985 census of Tampa Florida". Journal of the American Statistical Society 84 (406): 414–20.

2. Lauser, B., Johannsen, G., Caracciolo, C., et al (2008). Comparing Human and Automatic Thesaurus Mapping Approaches in the Agricultural Domain. In DC 2008 International Conference on Dublin Core and Metadata Applications, Berlin (Germany), 22-26 September 2008.

3. Van Hage, Willem (2008). Evaluating Ontology-Alignment Techniques.
See at http://www.few.vu.nl/~wrvhage/papers/wrvh_thesis_20080724.pdf

4. COMA++: http://dbs.uni-leipzig.de/de/Research/coma.html

5. Silk: http://www4.wiwiss.fu-berlin.de/bizer/silk/

6. FALCON-AO: http://ws.nju.edu.cn/falcon-ao/index.jsp

7. Simple Knowledge Organization System (SKOW): http://www.w3.org/2004/02/skos/

8. Jérôme David, Jérôme Euzenat, François Scharffe, Cássia Trojahn dos Santos, The Alignment API 4.0, *Semantic web journal* 2(1):3-10, 2011
9. Sesame: http://www.openrdf.org/