Framework for Matching and Linking Large Ontologies

Kow Weng Onn[1], Michelle Lim Sien Niu[1], Gudrun Johannsen[2], Johannes Keizer[2], Dickson Lukose[1],

[1]MIMOS Berhad
Technology Park Malaysia, Kuala Lumpur, Malaysia 57000

{kwonn | michelle.lim | dickson.lukose}@mimos.my

[2]The Food and Agricultural Organization of UN (FAO), Rome, Italy

{Gudrun.Johannsen| Johannes.Keizer}@fao.org

Abstract. The Linked Open Data (LOD) initiative is the first large-scale attempt at realizing the Semantic Web vision of Tim Berners-Lee. As of September 2011, there are almost 300 knowledge bases linked together and available for public access through both web browsers as well as semantic applications. However, the task of building the links between knowledge bases is still a laborious manual task and as the LOD cloud continues to grow rapidly, the task becomes more daunting. This paper looks at the AGROVOC thesaurus developed and maintained by the United Nations Food and Agriculture Organization (UN FAO), and how a framework can be developed to semi-automatically discover links between AGROVOC and other knowledge bases on the LOD.

Keywords: Linked Open Data, Ontology Matching, AGROVOC Thesaurus

1 Introduction

The Linked Open Data (LOD) initiative is the first large-scale attempt at fulfilling the Semantic Web vision of Tim Berners-Lee [25].However, despite having a large number of datasets and knowledge in the form of RDF triples, there is still a lack of links between the various datasets. According to the State of the LOD Cloud website [24], the number of datasets as at September 2011 is 295 with over 31 billion triples. However, there are only 504 million links or we can look at it as only 1.6% of all triples are links. The inter-links are mainly in the Life Sciences and Publications domain. Also, a large number of knowledge bases (33.22%) link only to one other knowledge base, usually DBPedia.

In this paper, we outline the needs and challenges of linking a large ontology (AGROVOC) to other ontologies on the LOD. This would enable knowledge harvesting tools such as Melody[4] to retrieve resource from multiple LOD ontologies by following the created links. In Section 2, we will describe the AGROVOC thesaurus and its current links to other ontologies on the LOD. Section 3 describes the current method and process for finding and creating the links. It also highlights the problems and challenges of the existing process. Section 4 briefly describes an existing tool for mediating two ontologies that can be modified for use on the LOD. Section 5 explains in detail our framework for solving some of the issues described in Section 3. In Section 6, we will report on some early experiments that we have performed using the framework proposed in Section 5. We finally conclude this paper with some discussion and future plans.

2 AGROVOC Thesaurus

AGROVOC[1] is a multilingual agricultural thesaurus and is used world-wide by researchers, librarians, information managers and others, for indexing, retrieving, and organizing data in agricultural information systems. From a traditional thesaurus, AGROVOC has developed into a SKOS-XLS concept scheme[2], containing more than 40 000 concepts in 21 languages.

In 2011, AGROVOC was published as Linked Open Data (LOD) [3]. LOD enables structured data and metadata to be published and connected on the web so that it can be consumed by both human and machine. LOD is especially useful to the agriculture domain because there are already existing rich multi-lingual vocabularies such as AGROVOC that can be used as a base to link to [4].As of June 2012, AGROVOC has been aligned with fourteen vocabularies (EuroVoc [5], NALT [6], GEMET [7], LCSH [8], STW - Thesaurus for Economics [9], RAMEAU [10], TheSoz [11], DBpedia [12], DDC [13],Geopolitical Ontology [14], SWD [15], GeoNames[16], ASFA[17] and FAO Biotechnology Glossary[18]) and it is not only linked to the agricultural domain but also related domains such as environment, economics, social sciences, geography, aquatic science and biotechnology. This has made AGROVOC the first and the largest LOD in the agriculture domain.

Figure 1 depicts an overview of the fourteen data resources that AGROVOC is linked to on the Linked Open Data [19].

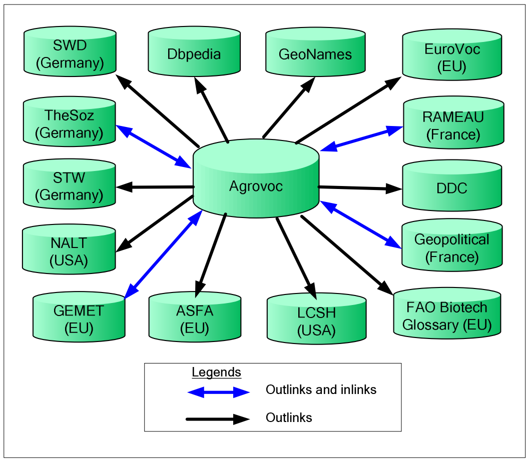

Fig.1. Agriculture Linked Open Data

© 2011 MIMOS Berhad. All Rights Reserved

Table1 provides the listings of the fourteen aligned vocabularies with AGROVOC: name of the aligned vocabulary (column 1), the domain of the vocabulary (column 2), languages available in the vocabulary (column 3) and the number of AGROVOC out-links aligned with its respective vocabulary (column 4). See [23] for some of the previous figures updated prior to Table 1.

Table 1. AGROVOC out-links on LOD (as of June 2012)

Vocabulary	Domain	Language	Out-links (from AGROVOC)
EuroVoc	General EU	EN, ES, DE, FR, etc. (24 languages)	1,297
GEMET	Environment	EN, ES, DE, FR, etc. (29 languages)	1,191
LCSH	General	EN	1,093
NALT	Agriculture	EN, ES	13,390
STW	Economy	EN, DE	1,136
TheSoz	Social Science	EN, DE	846

RAMEAU	General	FR	686
DBpedia	General	EN, ES, DE, FR, etc. (97 languages)	993
DDC	General	EN, ES, DE, FR, etc. (12 languages)	409
Geopolitical Ontology	Geopolitical	AR, ZH, FR, EN, ES, RU, IT	253
SWD	General	DE	5,965
GeoNames	Geographical database	67 languages	212
ASFA Thesaurus	Aquatic Sciences	EN, FR, ES	1,812
FAO Biotechnology Glossary	Biotechnology	AR, ZH, EN, FR, RU, ES, PL, SR, VI	791
Total			**30,074**

3 Linking AGROVOC to the rest of the LOD

Linking AGROVOC to other vocabularies allows access to document repositories and other agricultural data which are indexed, classified or organized by means of the interlinked metadata sets.This is the way to achieve interoperability of data in the agricultural domain. Alignments are not only done with English-to-English labels but also with other languages such as the French subject heading RAMEAU which only has concepts in French language. It has been mapped with the AGROVOC concept in French labels. These kinds of advantages enable the multilingual concept scheme to join the different resources and publish them as Linked Open Data (LOD)[20].

As an example, AGROVOC Linked Data is used as a backbone in OpenAgris[21], a web application to aggregate information from different Web sources with more than 60 million triples.

As AGROVOC is currently linked to fourteen data sets on LOD, it sets the milestone as being the first and the largest LOD in the agriculture domain.The process of aligning AGROVOC with six different vocabularies is described by Morshed et al. (2011)[22]. Only ‘preferred’ labels were considered, and the alignment was limited to ‘exact match’ links.The candidate matches obtained by running the matcher, were manually evaluated by a highly experienced domain expert from FAO, using the following criteria:

1. Check if there are non-preferred terms (alternative labels in SKOS terminology) associated with the candidate match term in order to clarify the meaning. If this not the case, then
2. Compare the matching term with other languages in common between the two thesauri, if available. AGROVOC and NALT (National Agricultural Library Thesaurus (of the U.S.A.), for example, have in common Spanish and English.
3. Take a look at the concept hierarchy, i.e. mainly parent concepts, and
4. Examine definitions or scope notes of mapped concepts, if available, to verify the correctness of exact matches.

The process is further illustrated in Figure 2 below.

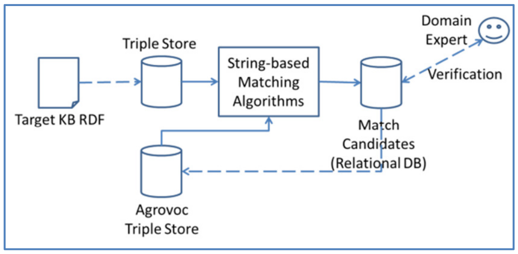

Fig. 2. Current process of aligning AGROVOC

There are a couple of issuesin this approach.

Firstly, the target ontology needs to be downloadable. This may not be available for all LOD ontologies or it might be too big. At present, only 117 of the 295 datasets (39.66%) are available as RDF dumps.

There is also the issue of the target ontology which may still be evolving and the version downloaded is not the latest. Links that are found and stored may not be correct or exist in the newer version of the ontology.

Secondly, there is the issue with time efficiency. By comparing every single possible pair, a lot of processing time is required. AGROVOC has over 30,000 concepts. Even matching that against ontology with only 100 concepts would require 3,000,000 comparisons. While this may not take a long time with simple String-based algorithms, more complex techniques will require hours or even days to process the ontologies.

Thirdly, even with reduction to candidate matches, evaluation by domain expert is still a manual and possibly tedious process. If matches are found for all AGROVOC concepts, then the expert must verify over 30,000 candidates to ensure that they are correct before they can be added as *exactMatch* in AGROVOC

Finally, the multilinguality of AGROVOC is not used to its full advantage. AGROVOC has labels for concepts in not just English but dozens of other languages. These labels can be used to automatically verify candidate matches. For example, the English label for the concept ‘Agroindustrial complexes’ in AGROVOC is found to have a high match with concept ‘Agroindustrial complex’ in STW. Both AGROVOC and STW have English and German labels. If the German labelsmatch as well, then there is a much higher possibility that the two concepts are indeed the same.

To solve the issues mentioned, we might implement the following:-

- Mediation through SPARQL endpoints. 201 (68.14%) datasets on the LOD provide SPARQL endpoints.
- Using crawlers to index LOD ontologies regularly
- Performing match as a search instead of an $m \times n$ comparison. Use simple algorithms such as String matching to filter out trivial matches before applying more complex algorithms to the remaining concepts. Parallel processing to make full use of computing resources
- Visualization and navigation tools aid the domain expert in making decisions. Crowdsourcing capability enables multiple domains experts to work on matching
- Indexing AGROVOC to multiple languages. If target ontology has labels in multiple languages, multiple searches can be used to get better match:
  i) Most confident: Labels match in more than 1 language. Can be automatically accepted without human intervention
  ii) Lesser confidence: Labels match in only 1 language. Give high score and may require human verification
  iii) Least confidence: Labels don’t match in multiple languages. Use another algorithm.

4 Semantic Mediation Tool

The Semantic Mediation Tool (SMT) [29] or Harmony, is an application for mediating between two ontologies that have been developed since 2009 by MIMOS and Know-Center. It consists of a Service-Oriented Architecture (SOA) back-end withmultiple similarity matching algorithms and a front-end with visualization and decision-support tools for verifying alignments discovered by the algorithms. The decision-support mechanism also enables the verification process to be split and parcelled out to various domain experts. This supports the crowd-sourcing methodology of decision-making by getting input from multiple experts instead of just one. Figure 3 depicts the overview of the Semantic Mediation Tool process flow whereby manual processes are represented by dashed arrow.

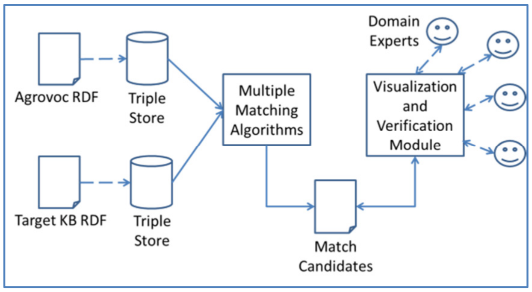

Fig. 3: Semantic Mediation Tool process flow. Dashed arrow shows manual processes

There are many existing algorithms for finding similarities between concepts. Euzenat and Shvaiko [27] in their seminal book on Ontology Matching have a very comprehensive list of the various algorithms. We are not going to list out the algorithms within this paper, but will point out that one advantage of the Harmony approach is that we implement the similarity matching algorithms as web-services and can plug-in many of the existing algorithms. We have also implemented several ways of integrating the results from multiple algorithms such as taking the average, the highest score and the lowest score.

5 Proposed Framework

We propose to modify the Semantic Mediation Tool to be able to handle ontologies residing on the LOD cloud. While no changes are done on the front-end interface, the modifications would be implementedon the following portions of the back-end (as shown in Figure 4).

i) Instead of loading the ontologies from a triple store, access would also be possible through SPARQL endpoints and RDF files on the web.
ii) An indexer would be built for indexing terms in the source ontology, AGROVOC in this case. We will be using Lucene[26], an open-source search and indexing software for this task.
iii) The index will then be used to search for candidates to apply the matching algorithms. This improves the performance by decreasing the number of comparisons that need to be made. Another benefit of using an indexing approach is that if we index multiple source ontologies, we can do a one-to-many matching instead of the current one-to-one matching. The ideal goal of course is to have the entire current LOD ontologies indexed so that links can be discovered whenever a new ontology needs to be added to the cloud.
iv) Instead of running similarity matching algorithms in parallel on all candidates, the system would first undergo a multi-stage filtering process with simple algorithms

at the beginning and then, applymore complex, time-consuming algorithms at the end. The hypothesis is that the ontology matching problem follows the "80:20" rule where 80% of the mappings can be found by simple algorithms taking up 20% of the processing time, while the other 20% of mappings can only be discovered through 80% of effort. In our case, we will be using the index itself as the first algorithm, a String-based algorithm next, followed by a taxonomy-based algorithm and finally a structural-based algorithm.

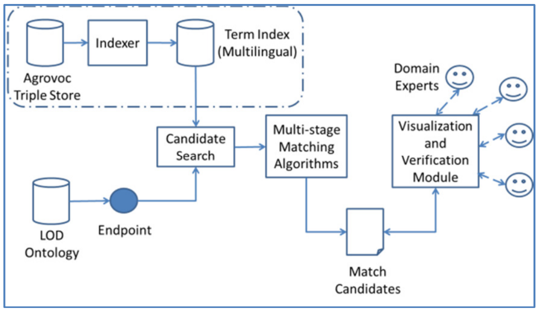
Indexer
Term Index (Multilingual)
Agrovoc Triple Store
Domain Experts
Candidate Search
Multi-stage Matching Algorithms
Visualization and Verification Module
LOD Ontology
Endpoint
Match Candidates

Fig. 4 – Architecture of proposed framework. The indexing of AGROVOC needs to be only done once

6 Experiments

The implementation of the proposed framework is being done in phases. For the first phase, we are looking at the feasibility of the "mediation as a search" approach. This will be an additional algorithm that will serve as the first simple filter that will reduce the number of matches that need to be made later. We then sequence a String matching algorithm based on Jaro-Winkler [28] after it to see if we can get even better results.

One of the challenges of testing any ontology matching algorithm is the lack of a benchmark, or "gold standard" for comparison. This is because to create a gold standard, especially for large ontology matching tasks, require a lot of human effort. In this case, we are using the existing links between AGROVOC and the STW Thesaurus for economics as the benchmark.

Preliminary experiments have given some encouraging results. The following experimental set-up was used

- The prototype was run on a quad-core Intel I7 2.7 GHz notebook with 4 GB RAM
- The source and target thesauri were AGROVOC and the STW Thesaurus of Economics. Both were converted into AllegroGraph Triple Stores for performance purposes.
- English preferred labels were used for indexing and matching
- Lucene version 3.5 was used to index the STW concepts
- A threshold is used to only consider search results that are above a certain Lucene score
- Results were compared with the existing links between AGROVOC and STW concepts

Three separate experiments were conducted at each threshold. In the first, just a simple Index and Match were conducted. Based on the findings of the previous research [22] into mapping AGROVOC, one of the issues was that AGROVOC uses plural terms such as "health foods" whereas STW uses the singular "health food" as the label.

The second experiment included a simple stemmer that converts plural nouns into singular nouns. For example, "Agroindustrial complexes" is stemmed to "Agroindustrial complex" which can then be matched with the same concept in STW.

The third experiment enhances the system further by using a Jaro-Winkler string matching algorithm to verify matches that are above the threshold but below a secondary threshold. This greatly improves the precision by removing mismatches found by setting the threshold too low.

In the current mapping, there are 1136 links between AGROVOC and STW. The precision and recall are calculated as follows.

Precision = (Correct mappings found)/(Total mappings found). Precision measures how high a percentage of accurate results are returned as having a lot of bad results can confuse the user as it "pollutes" the results returned with a lot of garbage.

Recall = (Correct mappings found)/1136. Recall, on the other hand, measures how good the algorithm is at finding matches. There is no point in having a high precision if only a small percentage of the actual mappings are found. A high recall means that the algorithm can find mappings as well as a human can.

6.1 Experimental Results

Table 2 – Results with primary threshold = 4.0, secondary threshold = 6.0

	Mappings found	Correct Mappings	Precision	Recall
Experiment 1	1587	1005	0.633	0.885
Experiment 2	1624	1025	0.631	0.902
Experiment 3	1389	1062	0.765	0.935

Table 3 – Results with threshold = 5.0, secondary threshold = 6.0

	Mappings found	Correct Mappings	Precision	Recall
Experiment 1	1420	1006	0.708	0.886
Experiment 2	1445	1021	0.707	0.899
Experiment 3	1293	1037	0.802	0.913

Table 4 – Results with threshold = 6.0, secondary threshold = 7.0

	Mappings found	Correct Mappings	Precision	Recall
Experiment 1	986	847	0.859	0.746
Experiment 2	1005	857	0.852	0.754
Experiment 3	996	855	0.858	0.753

At first glance, the minor additions of the stemmer and the Jaro-Winkler validation do seem to give a small improvement. The improvement is more appreciable at the lower threshold to the point that the high recall score may mean that it is a good compromise to have a lower precision in return for the higher number of matches found.

From the 3 tables above, it can be seen that both the precision and recall are affected by the Lucene score threshold. The higher the Lucene score threshold, the higher the precision but lower the recall. This matches the intuitive thought that setting the threshold bar higher would reduce the occurrence of erroneous matches, thus increasing the precision. However by enforcing a high standard, the algorithm also misses out some correct matches.

In light of the fact that our system will be a semi-automatic one with human validation, setting a low threshold to increase the recall would be the better option. After all, there is no use in having a human to validate the mappings if the system doesn’t find and present them to the user. However, as having a low threshold reduces the precision, we have to make sure that the user’s expectations are tempered and they realize that the system will sometimes make wrong alignments that they will have to correct manually.

7 Discussion and Future Work

As can be seen from the results, our simple prototype gives a good showing compared to existing links. It is also very fast, with execution times below 200 seconds including the time needed for indexing. Of course for our experiments we accessed the source and target ontologies from a triple store. When getting concepts through SPARQL endpoints, some extra overhead will be incurred due to issues beyond our control such as network latency and the speed of the triple-stores accessed by the endpoint.

One of the biggest issues we face from the experiments is the setting of the threshold based on the Lucene score. As this score is dependent on the index, the threshold that would give the best results in terms of precision and recall would differ for each target ontology we want to link to. This might involve a lot of manual tweaking. Further research into automatically proposing a good threshold would be helpful.

The precision score is also affected by matches that the system found that do not exist in AGROVOC at the moment but at cursory glance seem to be correct otherwise. For example, the system found that "Body weight" (http://aims.fao.org/aos/agrovoc/c_15846) in AGROVOC should be an exact match with "Body weight" (http://zbw.eu/stw/descriptor/28952-5) in STW. The fact that the German labels (Körpergewicht) also match means a high possibility that a link should exist.

We will be conducting more experiments to compare results of the system with the existing links from AGROVOC to the other ontologies such as EUROVOC, NALT and LCSH.

We would also like to try and map AGROVOC to another LOD ontology where no links exist currently to gauge how useful the tool will be. However, this will require a domain expert to verify and validate the mapping candidates discovered.

We have only started basic exploration into using multiple languages for finding mappings and will continue to modify the algorithms to take into account every piece of information that might assist in getting better results.

Finally, we will be modifying the existing Harmony interface to provide visualization and decision-making support for users to view and verify the candidate mappings.

References

1. _, (2012). AGROVOC. Food and Agriculture Organization of the United Nations (FAO), Rome, Italy.URL:http://aims.fao.org/standards/agrovoc/about (last visited June 2012).
2. _, (2012). AGROVOC Thesaurus Concept Scheme. Food and Agriculture Organization of the United Nations (FAO), Rome, Italy. URL:http://aims.fao.org/standards/agrovoc/concept-scheme (last visited June 2012).
3. _, (2012). AGROVOC Linked Open Data. Food and Agriculture Organization of the United Nations (FAO), Rome, Italy. URL:http://aims.fao.org/agrovoc/lod (last visited June 2012).
4. Lim, S. N., Kow, W. O., Lim, Y. S., Lukose, D.: Agriculture Linked Open Data, The Joint International Symposium on Natural Language Processing and Agricultural Ontology Service 2011 (*SNLP-AOS* 2011), Bangkok (2011)
5. EuroVoc (2011). EuroVoc: Multilingual Thesaurus of the European Union. URL: http://eurovoc.europa.eu/drupal/ (last visited August 2011).
6. NALT (2011). Agriculture Thesaurus and Glossary, National Agriculture Library, United States Department of Agriculture. URL: http://agclass.nal.usda.gov/agt.shtml (last visited April 2011).
7. GEMET (2011). General Multilingual Environmental Thesaurus, European Topic Center on Catalogue of Data Sources, European Environmental Agency. URL: http://www.eionet.europa.eu/gemet (last visited April 2011).
8. LCSH (2011). Library of Congress Subject Headings, United States Library of Congress, United States of America. URL: http://www.loc.gov/catdir/cpso/lcco/ (last visited April 2011).
9. STW (2011). STW Thesaurus for Economics, Leibniz Information Center for Economics, Kiel, Germany. URL: http://zbw.eu/stw/versions/latest/about (last visited April 2011).
10. RAMEAU (2011). Répertoired'autorité-matièreencyclopédiqueetalphabétiqueunifié, National Library of France, France. URL: http://rameau.bnf.fr/ (last visited April 2011).
11. TheSoz (2011). Thesaurus Sozialwissenschaften (TheSoz) at GESIS Linked Data Prototype. URL: http://lod.gesis.org/thesoz/ (last visited June 2012).
12. DBpedia (2011). About DBpedia. URL: http://dbpedia.org/page/DBpedia (last visited August 2011).
13. DDC (2011). Dewey Decimal Classification / Linked Data. URL: http://dewey.info/(last visited June 2012).
14. Geopolitical ontology (2011). FAO Country Profiles. Geopolitical information. URL: http://www.fao.org/countryprofiles/geoinfo.asp (last visited June 2012).
15. SWD (Schlagwortnormdatei) (2012). URL:http://www.ib.hu-berlin.de/texte/hausarbeiten/capellaro/swd-capellaro.htm (last visited June 2012).
16. GeoNames (2012). URL: http://www.geonames.org/ (last visited June 2012).
17. ASFA Thesaurus (2012). Aquatic Sciences and Fisheries Abstracts. URL: http://www.fao.org/fishery/asfa/en (last visited June 2012).
18. FAO Biotechnology Glossary (2012). FAO Glossary of Biotechnology for Food and Agriculture. URL: http://www.fao.org/biotech/biotech-glossary/en/ (last visited June 2012).
19. AGROPEDIA (2012). URL: http://agropedia.iitk.ac.in/ (last visited June 2012).
20. Lim, S. N., Johannsen, G., Morshed, A., Rajbhandari, S., Keizer, J., Kiran, L., Thunkijjanukij, A., Selan, N. E.: Multilinguality in AGROVOC Concept Scheme: Challenges and Experiences:URL:http://www.fao.org/docrep/article/am813e.pdf(last visited June 2012).
21. _, (2011). OpenAGRIS. Food and Agriculture Organization of the United Nations (FAO), Rome, Italy. URL:http://aims.fao.org/openagris (last visited June 2012).
22. Morshed, A., Caracciolo, C., Johannsen, G., Kizer, J. (2011): Thesaurus alignment for Linked Data publishing. International Conference on Dublin Core and Metadata Applications 2011, URL:http://www.fao.org/docrep/015/an895e/an895e00.pdf (last visited June 2012).
23. Caracciolo C., Stellato, A., Morshed, A., Johannsen, G., Rajbahndari, S., Jaques Y. and Keizer, J.: The AGROVOC Linked Dataset. URL: http://semantic-web-journal.org/sites/default/files/swj274.pdf (last visited June 2012).
24. Bizer, C., Jentzsch, A., Cyganiak, R.: State of the LOD Cloud (2011). URL: http://www4.wiwiss.fu-berlin.de/lodcloud/state/ (last visited June 2012)
25. Berners-Lee,Tim. Linked Data - Design Issues, 2006. URL : http://www.w3.org/DesignIssues/LinkedData.html (last visited June 2012)
26. _, (2012). Apache Lucene. URL:http://lucene.apache.org/ (last visited June 2012).
27. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer-Verlag (2007)

28. Winkler, William: The state of record linkage and current research problems, Technical Report 99/04, Statistics of Income Division, Internal Revenue Service Publication (1999)
29. Kow, W. O., Sabol, V., Granitzer, M., Kienreich, W., Lukose, D.: A visual SOA-based ontology alignment tool. Ontology Matching Workshop, 10th International Semantic Web Conference (2011)