Thesaurus Alignment for Linked Data Publishing

Ahsan Morshed
FAO, Italy
ahsan.morshed@fao.org

Caterina Caracciolo
FAO, Italy
caterina.caracciolo@fao.org

Gudrun Johannsen
FAO, Italy
gudrun.johannsen@fao.org

Johannes Keizer
FAO, Italy
johannes.keizer@fao.org

Abstract

As part of the publication of the AGROVOC thesaurus as Linked Data (LD), AGROVOC is now mapped with six well-known thesauri in the agricultural domain, i.e., EUROVOC, NALT, GEMET, STW, LCSH, RAMEAU. To find matching candidates, known matching algorithms discussed in the literature and available from public API were used. Results were evaluated by a domain expert, and almost total precision obtained. The candidate matches that were confirmed have already been added to the LD version of AGROVOC. Moreover, the owners of two of the thesauri mapped with AGROVOC have included in their data the mapping we identified. From this work, we conclude that we achieved our goal to enhance the Linked Data version of AGROVOC with reliable links to other thesauri, following a procedure that is fully replicable.

Keywords: Ontology Mapping, SKOS, Thesauri, Vocabularies, AGROVOC, Linked Data.

1. Introduction

The development of a Web of Data, built by applying Linked Data (LD) (Berners-Lee, 2011) (Heath, 2011) principles and using Semantic Web technologies, is gaining great attention in the academic as well as the industrial world. This is the frontier of data integration and sharing. In a web where each piece of data is published by means of standard technologies and data formats, and where each piece of data can be univocally named and located, data integration (understood as the possibility of programmatically accessing data residing in different sources) is perceived to be closer now than ever before. More and more data sets are now published as Linked Data and certainly more are going to be published soon: the cloud is growing, and so are the links inside. The central notions of LD are dereferenceable identifiers of resources (URIs), machine readable data in RDF/XML format, HTTP protocol, links to move from one resource to another.

For the bibliographic and librarian world, Linked Data offers the technology and the social attention needed to publish and interlink metadata sets: the advantage is the access to all documents and resources indexed/classified/organized by means of the interlinked metadata sets. If, for example, a term in the AGROVOC thesaurus is linked with a term in the GEMET thesaurus, all documents indexed by the same term in the document repositories related to AGROVOC and GEMET are also potentially linked. Using appropriate applications, information queries can be submitted against both repositories, and data results presented (and processed) to the user in a unified way. For this reason, many thesauri are adopting the Linked Data approach to data publishing. In this paper we present our work on aligning AGROVOC with six relevant thesauri, in order to publish AGROVOC as Linked Data.

The process of linking data sets may be very challenging, due to likely differences in formats, structure, semantics, and concept labels with different languages. Also, minor differences in spelling adopted and other formal conventions may prove problematic for thesaurus alignment.

The best-known related initiative is OAEI[1] that started in 2004. However up to now little attention has been dedicated to aligning thesauri, in particular for the purpose of LD publishing.

SKOS is now the format for publishing thesauri over the web, as it is a RDF vocabulary specific to the terminology and structure of thesauri. In the SKOS modeling, preferred and non-preferred terms are all labels of the same concept, and this applies to all languages available (Isaac et al, 209). In other words, in the SKOS modeling, a thesaurus is transformed into a set of concepts hierarchically organized by the usual BT/NT (broader/narrower) relationships, and all terms in the thesaurus in all languages are considered as labels of the same concept.

Our goal is to enrich the SKOS/Linked Data version of AGROVOC with appropriate links to other thesauri. The procedure adopted has to be replicable, and the resulting data has to be reliable enough to be published as part of the AGROVOC Linked Data. In this first phase of our work, we limited ourselves to *exact match links.* In SKOS terminology, two concepts are stated to be *exact match* if they can be used interchangeably in information retrieval applications (which can be taken as an operational approximation of having the same *meaning*). One issue we needed to pay special attention to is the fact that AGROVOC and many other thesauri are multilingual resources, where each concept may be "named" in as many as one or more than a dozen languages.

The remainder of this paper is organized as follows: In section 2 we introduce previous work related to resource alignment. In section 3 we introduce AGROVOC and the thesauri to which it was aligned. In section 4 we describe our approach to thesaurus mapping. We present and discuss the results obtained in section 5, and finally, in section 6 we draw some conclusions and hint at future work.

2. Related Studies

The problem of *matching* or *aligning* (Noy, 2004) (Euzenat et al., 2007) information resources such as XML schemas, database schemas, ontologies and the like, has received much attention as a pre-requisite to data exchange. Since 2004, the Ontology Alignment Evaluation Initiative is the international event to compare on a common benchmark the state of the art matching systems

A number of matching systems (Do et al., 2003) have been tested within the OAEI, most notably COMA++, RiMOM, FALCON-AO (Jian et al., 2005), and S-match (Giunchiglia, 2007), that use different approaches to computing string similarity. Systems like COMA++, RiMOM, and FALCON-AO analyze the input schema and reference mappings, and include rules for mapping. All these systems, however, use the OWL format and are focused on monolingual ontologies. Matching techniques may take into account only the *strings* representing the entities to match: in a string based approach, "book" and "booklet" would be taken as similar to some degree (exact value of similarity depends on the measure adopted), while "book and "volume" in no case would be considered as similar. Some approaches may use external resources to introduce a notion of *meaning* (in this case, depending on the approach taken, "book" and "volume" could be taken as similar). S-match uses WordNet as a background knowledge repository. Given that WordNet has general domain coverage, the tool provides good results in general domain, but performs poorly in specific domains like agriculture, forestry, etc. Finally, other approaches may also take into account other type of information, such as hierarchical information data structure when available (Aleksovski, 2006).

Relatively little experience is available concerning the alignment of thesauri for the purpose of Linked Data publication. Currently, STW, GEMET, LCSH and RAMEAU (see sec. 3 for an introduction to the thesauri mentioned) are available as Linked Data. In many cases, links are established manually, which we consider a bottleneck in the process of publishing Linked Data. Therefore, we went for a combination of candidate matches automatically identified and then

[1] Ontology Alignment Evaluation Initiative, http://oaei.ontologymatching.org

manually assessed, and looked at aligning techniques based on string similarity. These types of techniques seemed appropriate given that we deal with thesauri (i.e., standard controlled vocabularies), and we addressed the problem of aligning thesauri for the first time. In the following we mention some of the best-known string-based similarity measures, which are also those we used in our work (see sec. 4).

Some string equality measures take into account the number and proximity of the common characters between two strings (Cohen et al., 2003). Perhaps the most immediate way to compare two strings is to count the number of positions in which the two strings differ, as in the case of the Hamming distance (Hamming, 1950). Variations of this approach consider the common substrings between the string to compare, as in the case of the substring similarity, which looks at the longest common substring. A related notion of similarity is embodied by the n-gram similarity, where the number of common n-grams (i.e., sequences of n characters). This measure is efficient when only some characters are missing. Other commonly used measures are the edit distances, according to which the distance between two objects is the minimal cost of operations to be applied to one of the objects in order to obtain the other one. These measure are appropriate to measure strings that are spelling mistakes. The Levenshtein distance (Levenshtein, 1965) considers the operations of insertion, deletion and substitution, while the Needleman distance gives higher costs for insertion and deletion. Finally, The Jaro measure (Jaro, 1989) looks at common letters appearing the same positions in the two strings, and common letters that appear in different positions in the two strings (transposed). The Jaro-Winkler (Winkler, 1999) measure is a variation of the Jaro measure, that favors matching strings with longer prefixes. Another variation of the Jaro measure is the SMOA measure (Stoilos, 2005).

3. The Thesauri Aligned with AGROVOC

In this section we briefly introduce AGROVOC and the six thesauri to which it was mapped. We considered one thesaurus specific to agriculture (NALT), one specific to environment (GEMET), two general thesauri (LCSH, RAMEAU), one general but leaning to legal matters (EUROVOC), and STW, an economic thesaurus. While some of these resources are highly multilingual (EUROVOC, GEMET), others only cover a few languages (NALT, STW), while RAMEAU and LCSH are monolingual (French and English, respectively).

AGROVOC

AGROVOC[2] is managed by the Food and Agriculture Organization of the United Nations (FAO), and covers all its areas of interest, such as agriculture, forestry, fisheries, food and related domains. It is available in 21 languages, with an average of 40,000 terms per language. AGROVOC is available in SKOS (with close to 32,000 concepts), and published as Linked Data[3].

EUROVOC

EUROVOC[4] is managed by the European Union, and covers all areas of interest of the European Union, with special attention to parliamentary subjects. It is available in 24 languages. EUROVOC is available as a SKOS resource (Smedt, 2009), with close to 7,000 concepts.[5]

GEMET

GEMET[6], the GEneral Multilingual Environmental Thesaurus, covers the domain of environment, and it is available in 29 languages. It is published and managed by the European

[2] http://aims.fao.org/website/About/sub
[3] The HTML visualization of the Linked Data version of AGROVOC is available at http://aims.fao.org/website/Linked-Open-Data/sub
[4] http://eurovoc.europa.eu/
[5] The SPARQL endpoint for EUROVOC is: http://idi.fundacionctic.org/classifications_endpoint/eurovoc
[6] http://www.eionet.europa.eu/gemet

Environment Information and Observation Network. Its SKOS version consists of over 5,000 concepts, and it is also available as Linked Data[7].

LCSH

The LCSH[8] (Library of Congress Subject Headings) Thesaurus is the monolingual thesaurus (English) of subject headings, created for and maintained by the Library of Congress of the U.S.A. Its SKOS version consists of 30,000 concepts, and it is also available as Linked Data[9].

NALT

NALT[10], the National Agricultural Library Thesaurus, covers topics related to agriculture and is maintained by the National Agricultural Library of the U.S., USDA, and the Inter-American Institute for Cooperation on Agriculture (IICA) through the Orton Memorial Library, the Mexican Network of Agricultural Libraries (REMBA), as well as other Latin American agricultural institutions belonging to the Agriculture Information and Documentation Service of the Americas (SIDALC). It is available in two languages (English, Spanish). A SKOS version exists (consisting of some 30,000 concepts), but is not available as Linked Data.

RAMEAU

RAMEAU[11] (Répertoire d'Autorité-Matière Encyclopédique et Alphabétique Unifié, from French National Library) covers a variety of areas, such as geography, proper names, collective bodies and titles) and is available in French only. A SKOS version is available, which consists of about 150,000 concepts, and an experimental Linked Data service is available[12].

STW

STW[13] (Standard-Thesaurus Wirtschaft), Thesaurus for Economics is a bi-lingual (English, German) thesaurus of the German National Library of Economics. It covers law, sociology, politics, and geography. It is available as a SKOS resource, also published as Linked Data[14], and includes about 6,500 concepts (Neubert, 2009).

TABLE 1. Some figures about the thesauri aligned

Thesaurus	Topics	# Concepts	Languages available	Linked Data
AGROVOC	Agriculture, food, fishery, forestry..	31,956	EN, ES, DE, FR + 17 more	Yes
EUROVOC	General EU	6,779	EN, ES, DE, FR + 20 more	Yes
GEMET	Environment	5,298	EN, ES, DE, FR + 25 more	Yes
LCSH	General	30,784	EN	Yes
NALT	General	30,298	EN, ES	No, Only SKOS
RAMEAU	General	16,407	FR	Yes
STW	Economy	1,165	EN, DE	Yes

[7] http://svn.eionet.europa.eu/projects/Zope/wiki/GEMETLinkedData
[8] http://id.loc.gov/authorities/
[9] http://lcsubjects.org/
[10] http://agclass.nal.usda.gov/
[11] http://rameau.bnf.fr/
[12] http://www.cs.vu.nl/STITCH/rameau/
[13] http://zbw.eu/stw/versions/latest/about
[14] Experimental SPARQL endpoint at http://zbw.eu/beta/sparql at the time of writing this paper.

Table 1 summarizes some figures concerning the thesauri considered: the second column hints at the content of the resource, the third column reports the number of concepts available in the SKOS version. The forth column reports whether the thesaurus is also available as Linked Data.

4. Aligning Thesauri for Generating a Linked Data Version of AGROVOC

In this section, we describe the process followed to align AGROVOC with the selected thesauri, presented in the previous section. Figure 1 provides a schematic view of the process.

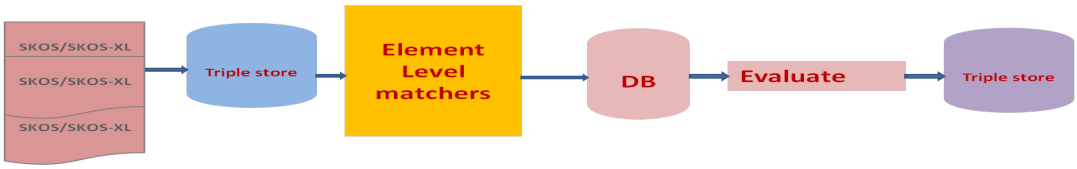

FIG.1. Matching process workflow

Since all thesauri considered are available as SKOS-RDF, we could load them all in a single local triple store (we used Sesame[15]). We considered the entire thesauri in all cases except in the case of RAMEAU, for which we selected only a set of concepts related to agriculture (amounting to some 10% of its 150 thousand concepts). Then, we considered all possible pairs of concepts, where the first concept in the pair comes from AGROVOC, and the second concept from one of the other thesauri. For each of the pair of concepts thus extracted, we computed various similarity values: we took one preferred label per concept (in the single language in common) and applied string similarity measures between those labels. Note that in this process only preferred labels in one language are considered because the matching methods do not support more than one language label at a time. The single language in common was English in all cases except for RAMEAU, which is a monolingual French thesaurus. Figure 2 recaps the thesauri selected and the languages used for alignment.

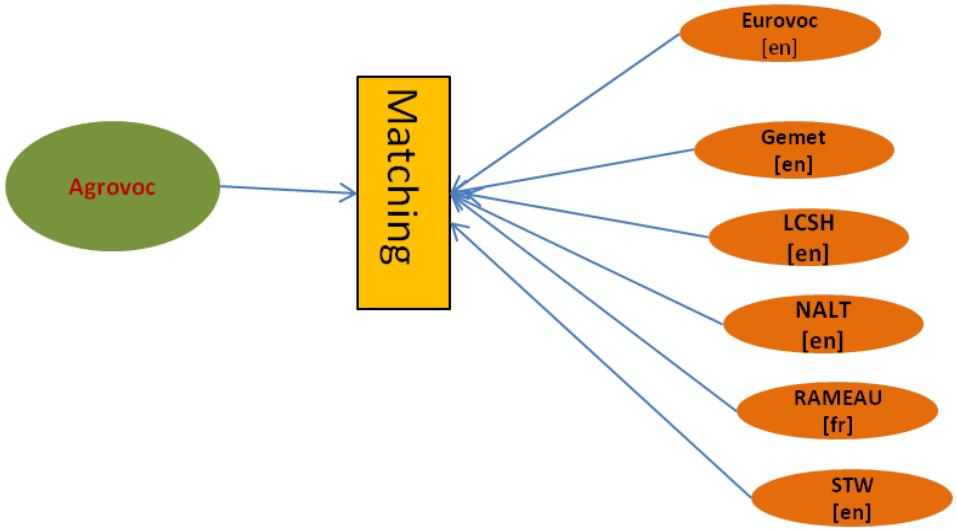

FIG.2. AGROVOC matching with other thesauri

We used a selection of the most common string similarity measures (those mentioned in sec. 2, last paragraph). The implementation used was the one made available through the Alignment API[16] (Euzenat, 2004),. In order to combine these similarity values into a single number, we

[15] http://www.openrdf.org/
[16] http://alignapi.gforge.inria.fr/

computed the arithmetic average of all similarity values, as the simplest way to combine several values, which seemed to us appropriate for a first attempt. Finally, an empirically identified threshold was applied to select the candidate matches to pass to human evaluation.

Once all candidate links were found, we produced data in a format suitable for manual evaluation by a domain expert. For this purpose candidate links were loaded into a relational database, and then exported as a spreadsheet. Candidate mappings that are confirmed by the domain expert are then loaded in the same triple store where the Linked Data version of AGROVOC is stored. This allows us to publish AGROVOC together with all its outbound links at the same time. We use Pubby[17] to serve as frontend of our data repository: AGROVOC is now published in the style of Linked Data publishing[18] (Figure 3 presents a screenshot of the human oriented visualization of AGROVOC as Linked Data: one can see the exact matches found for the concept Europe from AGROVOC).

When labels are totally identical, as in the case of *Animal protein*[19] from AGROVOC and *animal protein*[20] from EUROVOC, they are taken as exact match, with no further computation of similarity. When labels are not exactly identical, the similarity measures are applied, and the average of their value computed. For example, *Animal products*[21] from AGROVOC and *animal product*[22] from EUROVOC only differ by one letter, so they score high enough to pass the threshold, and can be considered as exact match.

FIG.3. Integrated mapping links

The storage of candidate mappings into a relational database is an ad-hoc solution to the problem of presenting data to our domain expert doing the evaluation (see next section for details about the evaluation). As our domain expert is very familiar with DB generated output, it was agreed

[17] http://www4.wiwiss.fu-berlin.de/pubby/
[18] http://aims.fao.org/standards/agrovoc/linked-open-data
[19] http://aims.fao.org/aos/agrovoc/c_439
[20] http://eurovoc.europa.eu/2845
[21] http://aims.fao.org/aos/agrovoc/c_438
[22] http://eurovoc.europa.eu/2737

this was the best way to present the results. The accepted matches were finally added to a triple store (where AGROVOC is stored) for enriching AGROVOC with those outbound links.

5. Results, Human Evaluation and Analysis

Table 2 summarizes the figures obtained by running our matcher, and the result of the human evaluation of the candidate matches found.

TABLE 2: Matching results and evaluation

		Manual evaluation		
Aligned thesauri	N. of candidate exact matches	N. of correct matches	N. of incorrect matches	Precision
AGROVOC-EUROVOC	1,321	1,298	23	98.26
AGROVOC-GEMET	1,240	1,190	50	95.97
AGROVOC-LCSH	1,166	1,095	71	93.90
AGROVOC-NALT	13,609	13,393	216	98.41
AGROVOC-STW	1,165	1,142	23	98.02
AGROVOC-RAMEAU	728	687	41	94.37
TOTAL	19,229	18,805	424	0.98

Candidate matches were evaluated by a highly experienced domain expert from FAO who has previously been involved in other thesaurus matching activities. The following guidelines were used:
To assess a candidate exact match suggested by the system, the following criteria need to be taken into consideration:

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, for example, have in common Spanish and English.*[23]
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 domain expert assessed all candidate matches, and we found that almost all candidate matches were confirmed (Table 2, last column). In total, the evaluation process required 40 working days. The high number of confirmed matches is due to the fact that thesauri express standard terminology of the domain they cover. Also, the fact that they agree so much in the preferred terms may be taken as a confirmation of their capacity to reflect common usage of words. Differences across thesauri are mainly due to the use of singular and plural. For example, English terms in AGROVOC are mainly plural while in EUROVOC, GEMET and STW terms appear in singular form. Similarly, French terms in AGROVOC are in singular form, while they appear as plural in RAMEAU. A clear source of incorrect matching however is when the two thesauri adopt different terms as preferred terms.

The few incorrect candidate matches may be classified as follows:

a) **Complete homonymy**. Consider for example: *flavouring* in AGROVOC (which refers to the action of adding flavour to a substance) and *flavouring* in EUROVOC (which refers

[23] Our evaluator is able to work in five languages (English, French, Spanish, German, and Italian).

to the substance added). The difference in meaning was found by consulting additional information available in the thesauri (as suggested in the evaluation guidelines): non-preferred term (AGROVOC: *flavour addition*; EUROVOC*: foodstuff with a flavouring effect*), BT (AGROVOC: *processing*; EUROVOC: *food additive*), and translations (AGROVOC: *aromatisation* (FR), *Aromatisieren* (DE), *Aromatizzazione* (IT); EUROVOC: *aromatizante* (ES), *Geschmacksstoff* (DE), and *sostanza aromatizzante* (II)).

b) **Near-homonymy**. Consider the case of *Calice* (AGROVOC) and *Calices* (RAMEAU). The meaning of *calice* as a concept in the botanical domain was verified by checking in AGROVOC the term hierarchy, BT *Périanthe,* and the non-preferred term *sépale,* while in RAMEAU the two parents (BT *Objets liturgiques*, and BT *Récipients à boire*) showed a completely different meaning.
c) **False friends: similar terms, but with different meaning**. Examples: *aviculture – apiculture*, *health – wealth*, *forest range – forest ranger*, *health care – health card*, *marché – marche*, or *Qualité de la viande – Qualité de la vie*.
d) **Other cases:** *collective farming* (AGROVOC: http://aims.fao.org/aos/agrovoc/c_1757) – *collective farm* (EUROVOC: http://eurovoc.europa.eu/983). Incorrectly, these two different concepts were mapped as 'exact match' although AGROVOC includes the exact matching term "*collective farms*" (http://aims.fao.org/aos/agrovoc/c_ 28845) which was not identified by the matcher.

6. Conclusion

In this paper we have presented our work on publishing the AGROVOC thesaurus as Linked Data, and in particular we described the process followed to provide outbound links from AGROVOC to six selected thesauri (EUROVOC, GEMET, LCSH, NALT, STW, and RAMEAU). We used simple string matching techniques, and reuse public implementations of them, to find *exact matches* (in SKOS terminology) between thesauri entries.

We only considered concept labels in the one language that AGROVOC has in common with each of the other candidate thesauri for mapping. The downloadable version of the thesauri in the SKOS format was used, but most of these thesauri are however already published as Linked Data (with the notable exception of NALT). Relatively few include links to other resources, in most cases they link to DBpedia (Auer et al., 2007) (e.g. STW, GEMET) with almost none to other thesauri.

During the evaluation process the concepts have been evaluated by checking not just English labels but also other language labels. Special attention has been given to the result evaluation and as a result our mapping links have been introduced into RAMEAU and GEMET thesauri.

To our knowledge, we performed the first massive alignment of thesauri for the purpose of publishing Linked Data, and we believe our experience may be useful to thesaurus managers and researchers in Linked Data alike. We found that simple string matching techniques are quite appropriate to provide candidate links in a Linked Data framework, as the human evaluation confirmed most of the matches found. Most of the steps in the process we followed were based on known algorithms and implementation, which makes us confident that the process may be repeated by other actors. On average, slightly more than a week was needed to complete, but we believe this time could be reduced working by applied some simple variations to the matching algorithms. For example, non-preferred terms could also be considered during the matching. Also, when applicable, the languages considered for matching could not be limited to one only. For example, if the languages in common are English, Spanish and French, we should not limit ourselves to look at the English labels only, but could consider labels in all three languages. This is in fact what the human evaluator did during the assessment phase (see guidelines in sec. 5). Some investigation could be devoted to phrase heuristics that may help thesaurus managers find the right balance between complexity of the matching algorithms used, and time dedicated to

manual assessment. Finally, a more standardized framework for human assessment, as opposed to the ad hoc created spreadsheet, could help thesauri managers in moving to Linked Data.

Acknowledgements

We wish to thank our colleagues in FAO, MIMOS[24] and ICRISAT[25] for their collaboration in various stages of this work. In particular, we are pleased to thank Yves Jaques, Lim Ying Sean, Sachit Rajbhandari, Prashanta Shrestha, Lavanya Neelam, and Armando Stellato. We also wish to thank Jérôme Euzenat, Stefan Jensen, Antoine Isaac, Søren Roug, Thomas Baker, and Mary Redahan.

References

Aleksovski, Z., Klein, M., ten Kate, W., van Harmelen, F. (2006). Matching unstructured vocabularies using a background ontology. In: Proceedings of Knowledge Engineering and Knowledge Management (EKAW).

Araujo, S., Houben, G.-J., Schwabe, D., Hidders, J. (2011). Fusion – Visually Exploring and Eliciting Relationships in Linked Data In Proceedings of ISWC.

van Assem, M., Malaisé, V., Miles, A., Schreiber, G. A. (2006) Method to Convert Thesauri to SKOS. In The Semantic Web: Research and Applications. 95-109.

Auer, S., Bizer, C., Lehmann, J., Kobilarov, G., Cyganiak, R., Ives, Z. (2007) DBpedia: A Nucleus for a Web of Open Data. Aberer et al. (Eds.): The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, Springer .

Berners-Lee,T. Linked Data. http://www.w3.org/DesignIssues/LinkedData.html. [Retrieved April 10, 2011,]

Cohen, W. W.; Ravikumar, P. and Fienberg, S. E. (2003). A comparison of string distance metrics for name-matching tasks. In Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web (IIWeb-03).

Do, H.H., Melnik, S., Rahm, E. (2003). Comparison of schema matching evaluations, in: Proceedings of the International Workshop Web and Databases, Lecture Notes in Computer Science, vol. 2593, Springer, Berlin, 2003

Euzenat, J. (2004). An API for ontology alignment. In Proceedings of the 3rd International Semantic Web Conference, pages 698–7112, Hiroshima, Japan.

Euzenat, J., Shaiko, P. (2007). Ontology Matching.Berlin, New York:Springer-Verlag.

Giunchiglia, F., Yatskevich, M., Shvaiko, P. (2007). Semantic Matching: Algorithms and Implementation. Journal on Data Semantics IX .

Hamming, R. W. (1950) Error detecting and error correcting codes. Bell System Technical Journal. 29 (2): 147-160.

Heath. T, C. Bizer. (2011) Linked Data. Evolving the Web into a Global Data Space Morgan & Claypool.

Isaac, A. Summers, E. SKOS Simple Knowledge Organization System Primer. W3C Group Note, 2009. http://www.w3.org/TR/skos-primer/

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.

Jian, N., Hu, W., Cheng, G., Qu, Y. (2005). Falcon-AO: Aligning ontologies with falcon. In: K-Cap 2005 Workshop on Integrating Ontologies.

Levenshtein V.I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady **10**: 707–10.

Liang, A., M. Sini, Chang Chun, Li Sijing, Lu Wenline, He Chunpei and Keizer J. (2006). The mapping schema for Chinese Agricultural Thesaurus to AGROVOC. New review of hypermedia and multimedia, 12(1), 51-62. Retrieved from http://www.fao.org/docrep/008/af241e00.htm#Contents.

Lauser, B., Johannsen, G., Caracciolo, C., Keizer, J., van Hage, W., Mayr, P. (2008). Comparing human and automatic thesaurus mapping approaches in the agricultural domain. In: Proc. Int'l Conf. on Dublin Core and Metadata Applications, Universitätsverlag Göttingen.

Morshed, A, and Sini, M. (2009). Creating and Aligning Controlled vocabularies. Workshop on Advanced Technologies for Digital Libraries, Trento, Italy

[24] http://www.mimos.my/
[25] http://www.icrisat.org/

Neubert, J.: Bringing the "Thesaurus for Economics" on to the Web of Linked Data (2009). Proc. WWW Workshop on Linked Data on the Web, Madrid, Spain.

Noy, N. Semantic Integration (2004).A Survey of Ontology-Based Approaches. SIGMOND Record. Vol 33, No 4, December, Pages 65-70.

Smedt, J. De (2009): SKOS Extensions for the EUROVOC Thesaurus. In: 3rd Annual European Semantic Technology Conference.

Stoilos, G., Stamou, G., Kollias, S. (2005). A string metric for ontology alignment. In Proceedings of the 4th International Semantic Web Conference, pages 624–637, Berlin, Heidelberg. Springer-Verlag.

Trojahn, C., Quaresma, P., Vieira, R..(2010).An API for Multi-lingual Ontology Matching In Proceedings of LREC 2010.

Wang S., Isaac A., Schopman B., Schlobach S., and Meij, L.. (2009)..Matching multi-lingual subject vocabularies. In European Conference on Digital Libraries, pages 125–137

Winkler, W. (1999). The state of record linkage and current research problems. Technical report 99/04. Statistics of Income Division. Internal Revenue Service Publication.