Semantic Interoperability of Multilingual Language Resources by Automatic Mapping

P. Schmitz°, E. Francesconi°†, N. Hajlaoui°, B. Batouche°, A. Stellato+

°Publications Office of the European Union, Luxembourg
†Institute of Legal Information Theory and Techniques of CNR (ITTIG-CNR), Italy
+Dept. of Enterprise Engineering, University of Rome Tor Vergata, Rome, Italy
enrico.francesconi@publications.europa.eu
francesconi@ittig.cnr.it

Abstract. The PMKI project is an European Commission action aiming to create a public multilingual knowledge management infrastructure to support e-commerce solutions in a multilingual environment. Such infrastructure will consist in a set of tools able to create interoperability between multilingual classification systems (like thesauri) and other language resources, so that they can be easily accessible through a Web dissemination platform and reusable by small and medium-sized enterprises (SMEs), as well as by public administrations. In this paper the standards used to represent language resources and a methodology for automatic mapping between thesauri, based on an information retrieval framework, are presented.

Keywords: Semantic Interoperability, Language Resources, Information Retrieval, Semantic Mapping, Ontolex-Lemon

1 Introduction

Language barriers in the EU make the European market fragmented and decrease its economic potential. Particularly small and medium-sized enterprises (SMEs) are currently at a disadvantage compared to big companies to approach European and global markets for the high cost of providing multilingual services. The EU institutions aim to overcome language obstacles and increase cross-border e-commerce by building open multilingual tools and features free of charge.

For this reason the European Commission, through the ISA[2] program[1], launched a pilot project on creating a public multilingual knowledge management infrastructure (PMKI project). It is aimed to support e-commerce solutions in a multilingual environment by creating a set of tools, based on semantic Web technologies, to facilitate the development of multilingual facilities able to improve cross border accessibility of digital services and e-commerce solutions. In practical terms, overcoming language barriers on the Web means creating multilingual

[1] ISA2: Interoperability solutions for public administrations, businesses and citizens (https://ec.europa.eu/isa2/home_en)

lexicons (as vocabularies, thesauri, taxonomies, semantic networks), establishing links between concepts, as well as using them to support the accessibility of services and goods offered through the Internet.

This paper presents an overview of the PMKI project (Section 2), the standards adopted, based on the Ontolex-Lemon model for language resources (Section 3) and their interoperability (Section 4). A semi-automatic methodology for establishing semantic interoperability based on an information retrieval framework is then proposed (Sections 5, 6, 7). Finally, some experiments on the application of such methodology to a gold-standard data set of matching concepts (Sections 8, 9) and some conclusions (Section 10) are reported.

2 The PMKI project

PMKI aims to implement a proof-of-concept infrastructure able to expose and to harmonize internal (European Union institutional) and external multilingual lexicons aligning them in order to facilitate interoperability. Moreover, the project aims to create a governance structure for a possible public service, in order to extend systematically the infrastructure by integrating supplementary public multilingual taxonomies/terminologies. The need to have a public and multilingual platform with a role of hub able to collect and share language resources in standardized formats is essential to guarantee semantic interoperability of digital services. For instance, such platform is missing in CEF.AT[2], while it would provide an advantage for the development of machine translation systems, in particular for domain-specific ones (tender terminology, medical terminology, etc.). A platform like PMKI may represent a *one-stop-shop* harmonized multilingual lexicons repository at European level.

Complementary to the European Language Resource Coordination (ELRC[3]) action, which aims at identifying and gathering language and translation data, the PMKI platform aims firstly to harmonize multilingual language resources making them interoperable, then to integrate supplementary public multilingual taxonomies/terminologies in a standardized representation.

3 Standard representation of language resources

With the advent of the Semantic Web and Linked Open Data, a number of models have been proposed to enrich ontologies with information about how vocabulary elements have to be expressed in natural language. These include the Linguistic Watermark framework [1, 2], LexOnto [3], LingInfo [4], LIR [5], LexInfo [6] and Monnet lemon [7]. The lemon model envisions an open ecosystem in which ontologies and their lexicons co-exist, published as data on the Web.

[2] https://ec.europa.eu/digital-single-market/en/automated-translation

[3] ELRC: the European Language Resource Coordination action launched by the European Commission as part of the CEF.AT platform activities, to identify and gather language data across all 30 European countries participating in the CEF programme. More information can be found here: http://www.lr-coordination.eu/

In 2012, the OntoLex W3C Community Group[4] (OntoLex) was chartered to define an agreed specification informed by the aforementioned models, whose designers are all involved in the community group. The OntoLex Group published its final report[5] defining the OntoLex-Lemon model [8]: a suite of RDF vocabularies (called modules) for the representation of ontology lexicons in accordance with Semantic Web [9] best practices. The modules of OntoLex-Lemon cover aspects such as morphology, syntax-semantics mapping, variations, translation, and linguistic metadata. This rich linguistic characterization of ontologies is unattainable with widely deployed models on the Semantic Web (e.g. RDFS and SKOS-(XL) labels), and it enables a wide range of ontology-driven NLP applications (e.g. knowledge verbalization, semantic parsing, question answering...) [10]. Outside of its original scope, the OntoLex-Lemon model (and its predecessors) has been also used to represent and interlink lexicons, lexical-semantic resources and, in general, language resources in the Linguistic Linked Open Data (LLOD) cloud [11]. For such characteristics, OntoLex-Lemon has been adopted to represent language resources within the PMKI project. The OntoLex-Lemon

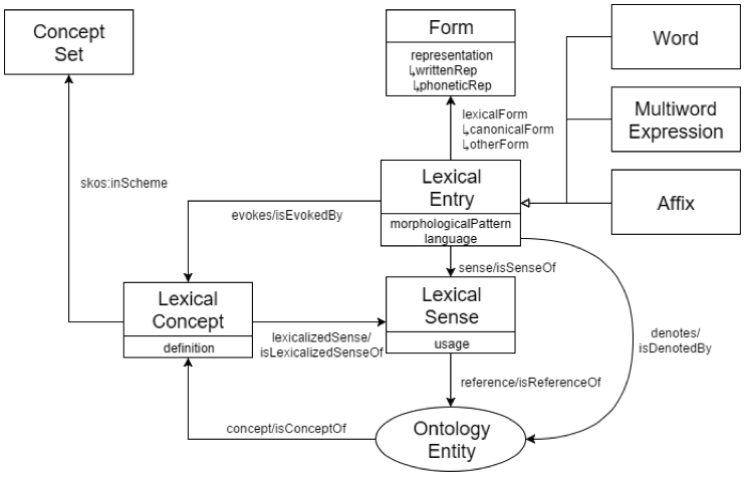


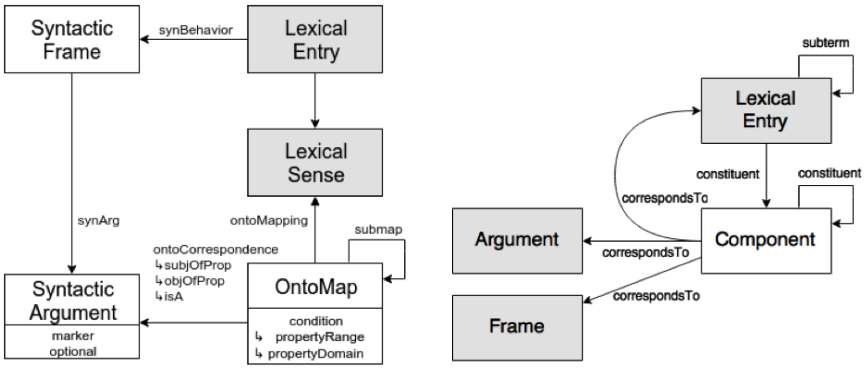
Fig. 1: The OntoLex Core data model

model is primarily based on the ideas found in Monnet lemon, which was already adopted by a number of lexica [12–14]. More specifically, OntoLex-Lemon consists of a number of vocabularies corresponding to different modules: core, synsem, decomp, vartrans, lime. The core module (Fig. 1) retains from Monnet lemon the separation between the lexical and the ontological layer (following [15] and [16]), where the ontology describes the semantics of the domain and

[4] Ontology-Lexica Community Group: https://www.w3.org/community/ontolex/

[5] Lexicon Model for Ontologies: Community Report, 10 May 2016: https://www.w3.org/2016/05/ontolex/ (last consulted: 03/04/2018)

the lexicon describes the morphology, syntax and pragmatics of the words used to express the domain in a language. A lexicon consists of lexical entries with a single syntactic class (part-of-speech) to which a number of forms are attached (e.g. the singular/plural forms of a noun), and each form has a number of representations (string forms), e.g. written or phonetic representation. While an entry can be linked directly to an entity in an ontology, usually the binding between them is realized by a lexical sense resource where pragmatic information such as domain or register of the connection may be recorded. Lexical concepts were introduced in the model to represent the "semantic pole of linguistic units, mentally instantiated abstractions which language users derive from conceptions". They are intended to represent abstractions in existing lexical resources such as synsets in wordnets.



(a) The syntax-semantics module (synsem)

(b) The decomposition module (decomp)

Fig. 2: The OntoLex modules

The synsem module (Fig. 2(a)) allows to associate a lexical entry with a syntactic frame (representing a stereotypical syntactic context for the entry), while an ontology mapping can be used to bind syntactic and semantic arguments together. The decomp module (Fig. 2(b)) is concerned with the decomposition of a lexical entry into its constituents (i.e. tokens). Components are instances of a dedicated class, which in turn correspond to lexical entries. This indirection allows recording inside a component information such as the fact that the entry "autonomo"es occurs with feminine gender inside "comunidad autonoma"es. We can also represent parse trees, by subdividing a component into its constituents. Because of the lack of space, we will not introduce vatrans and lime, but necessary information about them will be provided briefly later on. Additionally, [17] describes the design of (a candidate-release version of) LIME under the perspective of metadata-based discovery and exploitation of linguistic information in different tasks, including ontology mediation [18]. While Semantic Web

practitioners recognized the benefits of linguistic information, linguists in turn acknowledged that the adoption of Semantic Web technologies could benefit the publication and integration of language resources. This led to the formation of the Linguistic Linked Open Data (LLOD) cloud. There is thus a convergence of interests and results between these two communities. Unsurprisingly, recent discussions on OntoLex-Lemon were focused on improving its suitability to encode (legacy) language resources, departing from its original focus on ontology lexicons.

4 Semantic interoperability types in PMKI

PMKI aims to implement the following two types of semantic interoperability between language resources

- Semantic resources lexicalization (for example enriching thesaural concepts with lexical information)
- Conceptual mapping between semantic resources (for example identifying matching concepts in different thesauri)

In this work we have focused the attention on formalizing the conceptual mapping between semantic resources, in particular between thesauri, and to implement a semi-automatic procedure to establish mapping relation between concepts of two different thesauri.

5 Thesaural mapping formal characterization

Conceptual mapping in PMKI is a problem of thesaural concepts alignment, having only thesaurus schema available (*Schema-based mapping* [20]). In this case thesaurus mapping is the problem of identifying the conceptual/semantic similarity between a descriptor (represented by a simple or complex term[6]) in a source thesaurus and candidate descriptors in a target thesaurus.

A vast literature exist in this field [19–21] combining different approaches, in [22] the schema-based Thesaurus Mapping ($\mathcal{TM}$) problem has been characterized as a problem of Information Retrieval ($\mathcal{IR}$): the aim is to find concepts in target thesaurus, better matching the semantics of a concept in a source thesaurus. The isomorphism between $\mathcal{TM}$ and $\mathcal{IR}$ ($\mathcal{TM} \equiv \mathcal{IR}$) can be established once we consider a source concept as a *query* of the $\mathcal{IR}$ problem, and a target concept as a *document* of the $\mathcal{IR}$ problem.

Therefore, the $\mathcal{TM}$ problem can be viewed and formalized, like the $\mathcal{IR}$ problem, as a 4-uple $\mathcal{TM} = [D, Q, F, R(\boldsymbol{q}, \boldsymbol{d})]$ [23] where:

1. D is the set possible representations (*logical views*) of a concept in a target thesaurus (a document to be retrieved in the $\mathcal{IR}$ problem);

[6] for example *Parliament* is a simple term, *President of the Republic* is a complex term.

2. Q is the set of the possible representations (*logical views*) of a concept in a source thesaurus (a query in the $\mathcal{IR}$ problem);
3. F is the framework of concepts representation in source and target thesauri;
4. $R(\boldsymbol{q}, \boldsymbol{d})$ is a ranking function, which associates a real number with $(\boldsymbol{q}, \boldsymbol{d})$ where $\boldsymbol{q} \in Q$, $\boldsymbol{d} \in D$, giving an order of relevance to the concepts in a target thesaurus with respect to a concept of a source thesaurus.

In this framework the implementation of a thesaurus mapping procedure is represented by the instantiation of the previous 4 components.

6 Logical views (Q and D) of *descriptors* and matching framework (F)

Mapping between thesaural concepts is a process which aims at matching concept semantics rather than their lexical equivalences.

In PMKI thesaural concepts are represented by the SKOS model included in OntoLex-Lemon. In traditional thesauri, concepts are *descriptors* and *non-descriptors* represented by different terms (`skos:prefLabel` and `skos:altLabel`, according to SKOS) expressing the same meaning. More precisely, each meaning is expressed by one or more terms[7] in the same language (for instance 'pollution', 'contamination', 'discharge of pollutants'), as well as in different languages (for instance, the English term 'water' and the Italian term 'acqua', etc.). Moreover, each term can have more than one sense, i.e. it can express more than one concept. Therefore, to effectively map thesaural concepts, term (simple or complex) semantics has to be captured and represented.

In $\mathcal{IR}$ a query is usually constructed as a context (set of keywords) able to better represent the semantics of a query. Similarly, in $\mathcal{TM}$ the semantics of a thesaural concept is conveyed not only by its terms, but also by the context in which the concept is used, as well as by the relations with other concepts. In $\mathcal{TM}$ problem, Q, D and F are exactly aimed at identifying logical views and related framework for concept representations able to better capture the semantics of terms in source and target thesauri, as well as to measure their conceptual similarity.

In this work we propose to represent the semantics of a thesaural concept by a vector $\boldsymbol{d}$ of binary[8] entries composed by the term itself, relevant terms in its definition, in the alternative labels, as well as terms of directly related thesaural concepts (broader, narrower, related concepts).

Firstly a vocabulary of normalized terms from target thesaurus is constructed, where 'normalization' in this context means string pre-processing, in particular stopwords eliminations and word stemming/lemmatization procedures. In order to implement such pre-processing steps, the word stemming/lemmatization procedures provided by the java-based Elasticsearch libraries are used[9].

[7] Linguistic expressions by single or multi words.

[8] Statistics on terms to obtain weighted entries are not possible since document collections are not available (*schema-based thesaurus mapping*)

[9] https://www.elastic.co/guide/en/elasticsearch/guide/current/stemming.html

Being T the dimension of such vocabulary, both source and target concepts $\boldsymbol{d}$ are represented in a vector space of T-dimension ($\boldsymbol{d} = [x_1, x_2, \cdots, x_T]$); the entry x_i gives information on the presence/absence of the corresponding i^{th} vocabulary term among the terms characterizing the concept $\boldsymbol{d}$. In Fig. 3 a binary vector representation of a Eurovoc concept is sketched. In such representation the framework F is composed of T-dimensional vectorial space and linear algebra operations on vectors.

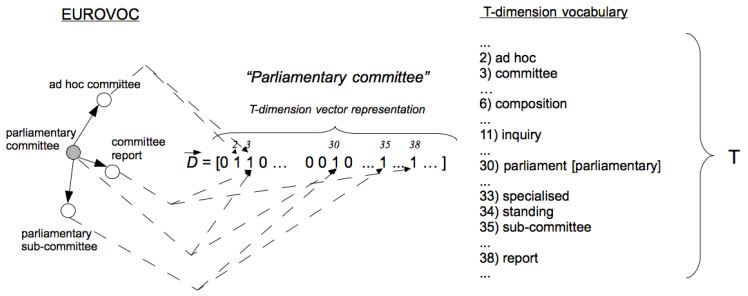


Fig. 3: T-dimension vectorial representation of a thesaural concepts $\boldsymbol{d}$.

7 The proposed ranking function (R)

Having represented the semantics of thesaural concepts as a binary vector, their similarity can be measured as the related binary vectors correlation, quantified, for instance, as the cosine of the angle between them

$$sim(\boldsymbol{q}, \boldsymbol{d}) = \frac{\boldsymbol{q} \times \boldsymbol{d}}{|\boldsymbol{q}| \cdot |\boldsymbol{d}|} \tag{1}$$

where $|\boldsymbol{q}|$ and $|\boldsymbol{d}|$ are the norms of the vectors representing concepts in source and target thesauri, respectively.

In order to classify a couple of concepts as an instance of the set of matching concepts (represented by one of the `skos:mappingRelation`), a heuristic threshold $Th \in [0, 1]$ is set as decision surface, so that:

$$\text{if } sim(\boldsymbol{q}, \boldsymbol{d}) \geq Th \Rightarrow (\boldsymbol{q}, \boldsymbol{d}) \in \texttt{skos:mappingRelation} \tag{2}$$

8 Interoperability assessment through a "gold standard"

In this work a thesaurus mapping case-study is proposed, including three thesauri of interest for the European Union institutions. The thesauri are EUROVOC, ECLAS, STW. EUROVOC is the main EU thesaurus containing a hierarchical structure with inter-lingual relations. It helps to coherently and effectively

manage, index, and search information of EU documentary collections, covering 21 fields. ECLAS is the European Commission Central Libraries thesaurus[10], covering 19 domains. STW[11] is the Thesaurus for Economics of the German National Library of Economics, bilingual thesaurus for representing and searching for economics-related content. It covers almost 6.000 subject headings in English and German and more than 20,000 synonyms.

The evaluation of the mapping procedure is based on a "gold standard" data set, namely an ideal collection of conceptual mappings expected by humans. To build the "gold standard" data set, an intellectual activity has been carried out by a group of experts dealing with EUROVOC as pivot thesaurus. The experts have established exact match relations between EUROVOC descriptors and the descriptors of ECLAS and STW, respectively. Specific guidelines have been given to the experts [24] to establish relations of type `skos:mappingRelation`, including `skos:exactMatch`, `skos:closeMatch`, `skos:narrowMatch`, `skos:broadMatch` relations. The complete "gold standard" dimension is reported in Tab. 1.

Thesauri	**Couples of matching concepts**
EUROVOC-ECLAS	4099
EUROVOC-STW	2959
Total number of matches	**7058**

Table 1: The "gold standard" of matching concepts

9 Experimental results

A set of experiments for thesaural conceptual mapping is carried out over the "gold standard". The experiments have been carried out using English as the pivot common language of all the three thesauri (anyway the approach is independent of the languages or of their combination, as long as they are the same for source and target resources). These experiments aimed at establishing the optimal similarity threshold, representing the best percentage of matching prediction, in terms of combination of *Precision* and *Recall* (*F-measure*), as well as *Accuracy*, of the cosine distance concept matching predictor. In Tab. 2 the detailed results of different experiment runs are reported, obtained by heuristically adapting the similarity threshold Th aiming at optimizing the automatic predictions quality over the gold-standard. The best results in terms of F-measure and Accuracy have been obtained using a similarity value threshold $Th = 0.3$, so that the mapping procedure is the best compromise between having a good level of Recall (so to include in the prediction the most part of actually matching concepts) and accuracy, while not decreasing too much in Precision.

[10] http://ec.europa.eu/eclas/
[11] http://zbw.eu/stw/

Similarity	**Eurovoc–Eclas**				**Eurovoc–STW**			
Threshold (Th)	P	R	F	A	P	R	F	A
0.60	98.89	64.41	80.17	82.70	99.50	52.91	69.09	75.87
0.50	98.32	77.24	86.51	87.50	98.46	72.21	83.32	85.26
0.40	90.43	87.16	88.77	88.55	94.29	85.54	89.70	89.99
0.30	**88.36**	**90.73**	**89.53**	**88.99**	**91.45**	**92.30**	**91.88**	**91.68**
0.20	86.53	92.47	89.40	88.63	88.02	96.02	91.84	91.31

Table 2: Precision (P), Recall (R), F-Measure (F) and Accuracy (A) results of the matching concepts prediction according to different values of similarity thresholds (Th)

10 Conclusions

The PMKI project aims to establish interoperability between language resources. While the OntoLex-Lemon model has been used to represent linguistic resources for the Semantic Web, two types of semantic interoperability are forseen: lexicalization and conceptual mapping. In this paper an approach for establishing automatic interoperability between language resources by semantic mapping of thesaural concepts has been presented. The preliminary results have shown satisfactory performance of the matching predictor. As future development we aim to implement a machine learning approach to set the matching decision surface on the basis of examples of matching concepts, as illustrated in [22].

References

1. M.T. Pazienza, A. Stellato, and A. Turbati. Linguistic watermark 3.0: an rdf framework and a software library for bridging language and ontologies in the semantic web. In *Proceedings of the 5th Workshop on Semantic Web Applications and Perspectives (SWAP2008)*, Rome, Italy, 2008.
2. A. Oltramari and A. Stellato. Enriching ontologies with linguistic content: an evaluation framework. In *Proceedings of OntoLex 2008*, Marrakech, Morocco, 2008.
3. P. Cimiano, P. Haase, and M. Herold et al. Lexonto: A model for ontology lexicons for ontology-based nlp. In *Proceedings of the OntoLex07 Workshop*, 2007.
4. P. Buitelaar, T. Declerck, and A. Frank et al. Linginfo: Design and applications of a model for the integration of linguistic information in ontologies. In *Proceedings of OntoLex06*, Genoa, Italy, 2006.
5. E. Montiel-Ponsoda, G. Aguado De Cea, and A. Gómez-Pérez et al. Enriching ontologies with multilingual information. *Natural Language Engineering*, XVII(3):283–309, 2011.
6. P. Cimiano, P. Buitelaar, and J. Mc Crae et al. Lexinfo: A declarative model for the lexicon-ontology interface,. *Web Semantics: Science, Services and Agents on the World Wide Web*, IX(1):29–51, 2011.
7. J. McCrae, G. Aguado-De-Cea, and P. Buitelaar et al. Interchanging lexical resources on the semantic web,. *Language Resources and Evaluation*, XLVI(4):701–719, 2012.

8. J. McCrae, J. Bosque-Gil, and J. Gracia et al. The ontolex-lemon model: Development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21, 2017.
9. T. Berners-Lee, J.A. Hendler, and O. Lassila. The semantic web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, CCLXXXIV(5):34–43, 2001.
10. P. Cimiano, C. Unger, and J. McCrae. Ontology-based interpretation of natural language. *Synthesis Lectures on Human Language Technologies*, VII(2):1–178, 2014.
11. C. Chiarcos, S. Nordhoff, and S. Hellmann, editors. *Linked Data in Linguistics*. Spinger, Berlin, Heidelberg, 2012.
12. L. Borin, D. Dannells, and M. Forsberg et al. Representing swedish lexical resources in rdf with lemon. In *Proceedings of the ISWC 2014 Posters & Demonstrations Track*, pages 329–332, Riva del Garda, Italy, 2014.
13. R. Navigli and S. Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, CXCIII:217–250, 2012.
14. J. Eckle-Kohler, J. Mccrae, and C. Chiarcos. Lemonuby–a large, interlinked, syntactically-rich lexical resource for ontologies. *Semantic Web*, VI(4):371–378, 2015.
15. P. Buitelaar. Ontology-based semantic lexicons: Mapping between terms and object descriptions. In C.R. Huang, N. Calzolari, and A. Gangemi et al., editors, *Ontology and the Lexicon*. Cambridge University Press, Cambrige, United Kingdom, 2010.
16. P. Cimiano, J. Mccrae, and P. Buitelaar et al. On the role of senses in the ontology-lexicon. In A. Oltramari, P. Vossen, L. Qin, and et al., editors, *New Trends of Research in Ontologies and Lexical Resources. Theory and Applications of Natural Language Processing*, pages 43–62. Springer, Berlin, Heidelberg, 2013.
17. M. Fiorelli, A. Stellato, and J.P. McCrae et al. Lime: The metadata module for ontolex. In F. Gandon, M. Sabou, and H. Sack et al., editors, *The Semantic Web. Latest Advances and New Domains*, LNCS, pages 321–336. Springer, 2015.
18. M. Fiorelli, M.T. Pazienza, and A. Stellato. A meta-data driven platform for semi-automatic configuration of ontology mediators. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14)*, Reykjavik, Iceland, 26-31 May 2014 2014.
19. P. Resnik. Disambiguating noun groupings with respect to wordnet senses. In *Natural Language Processing Using Very Large Corpora*, volume 11 of *Text, Speech and Language Technology*, pages 77–98, 1999.
20. P. Shvaiko and J. Euzenat. A survey of schema-based matching approaches. *Journal on Data Semantics*, 57, 2005.
21. Javier Lacasta, Javier Nogueras-Iso, Gilles Falquet, Jacques Teller, and F. Javier Zarazaga-Soria. Design and evaluation of a semantic enrichment process for bibliographic databases. *Data & Knowledge Engineering*, 88:94 – 107, 2013.
22. E. Francesconi G. Bartoloni. Sharing knowledge by conceptual mapping: the case of EU thesaural interoperability. In *Proceedings of the JURIX Conference*, pages 17–26. IOS Press, 2010.
23. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.
24. A. C. Liang and M. Sini. Mapping AGROVOC and the Chinese Agricultural Thesaurus: Definitions, tools, procedures. *New Review of Hypermedia and Multimedia*, 12(1):51–62, 2006.