Department of
Information Engineering
and Computer Science
DISI

UNIVERSITY
OF TRENTO - Italy

DISI - Via Sommarive 14 - 38123 Povo - Trento (Italy)
http://www.disi.unitn.it

SAM: A TOOL FOR THE SEMI-AUTOMATIC MAPPING AND ENRICHMENT OF ONTOLOGIES

Vincenzo Maltese, Bayzid Ashik Hossain

March 2012

Technical Report # DISI-12-007

Also: in proceedings of the 8th International Workshop on Ontology Content (OnToContent 2012).

SAM: A tool for the semi-automatic mapping and enrichment of ontologies

Vincenzo Maltese, Bayzid Ashik Hossain

DISI – University of Trento, Trento, Italy

Abstract. Ontologies are fundamental tools used with different purposes and with different modalities in different areas and communities. To guarantee the right level of quality, the most widely used ontologies are man-made. However, developing and maintaining them turns out to be extremely time-consuming. For this reason, there are approaches aiming at their automatic construction where ontologies are incrementally extended by extracting and integrating knowledge from existing sources. However, these approaches tend to reach an accuracy that, according to the application they need to serve, cannot be always considered satisfactory. Therefore, when a higher accuracy is necessary, manual or semi-automatic approaches are still preferable. In this paper we present a technique and a corresponding tool, that we called SAM (semi-automatic mapper), for the semi-automatic enrichment of an ontology through the mapping of an external source to the target ontology. As proved by our evaluation, the tool allows saving around 50% of the time required by purely manual approaches.

Keywords: Ontologies; mapping; semi-automatic enrichment

1 Introduction

Ontologies are used in different communities, for different purposes and with different modalities [2]. Many definitions of *ontology* have been provided. Studer et al. [3], by extending the famous definition by Gruber [4], define it as *a formal, explicit specification of a shared conceptualization*. The notion of conceptualization refers to an abstract model of how people theorize (the relevant part of) the world in terms of basic cognitive units called *concepts*. Concepts represent the intention, i.e. the set of properties that distinguish the concept from others, and summarize the extension, i.e. the set of objects having such properties. Concepts basically denote classes of objects. For instance, the medicine domain can be theorized in terms of doctors, patients, body parts, diseases, their symptoms and treatments used to cure or prevent them. Explicit specification means that the abstract model is made explicit by providing names and definitions for the concepts. In other words, the name and the definition of the concept provide a specification of its meaning in relation with other concepts. The specification is said to be formal when it is written in a language with formal syntax and formal semantics, i.e. in a logic-based language such as Description Logic [5]. The conceptualization is shared in the sense that it captures knowledge which is common to a community of people and therefore represents concretely the level of agreement reached in that community. By providing a common formal terminology (i.e. a vocabulary of terms) and understanding of a given domain of interest, ontologies allow

for automation (logical inference), support learning, reuse and favor interoperability across applications and people. When an ontology is populated with the instances of the classes, i.e. the individuals, it is called a *knowledge base*. In literature (see for instance [5]) the terms TBox and ABox are often used to denote what is known about the classes and about the individuals, respectively.

In order to guarantee the right level of quality, the most successful and widely used ontologies are man-made. We can mention for instance WordNet [6], Cyc [7], SUMO [8], Agrovoc[1] and UMLS[2]. The latter two are domain specific ontologies, in agriculture and medicine, respectively. However, maintaining them is extremely costly.

Attempts have been made to overcome this limitation by constructing ontologies automatically. One of the best examples in this direction is provided by YAGO [9], an ontology where the skeleton, constituted by WordNet, is progressively enriched with knowledge automatically extracted from Wikipedia[3]. This is done by mapping Wikipedia categories to WordNet synsets. Wikipedia categories can be seen as folders containing articles about individuals. For instance, the category *Italian scientists* contains an article about *Antonio Meucci*. WordNet synsets are groups of words which are synonyms, i.e. words with the same meaning, and corresponding definition. For instance, the synset containing the words *scientist* and *man of science* is defined as *a person with advanced knowledge of one or more sciences*. Basically, each category can be seen as a class of individuals that is mapped to a concept in the ontology; with the mapping, the ontology is enriched with knowledge coming from the *external source*. This mapping has a pretty high claimed accuracy of 90-95%.

As a matter of fact, computing the mapping between two ontologies is an essential step towards their integration [11]. Many projects have dealt with this problem. In the context of digital libraries this is a hot problem. We can mention for instance CARMEN[4], Renardus [13] and OCLC initiatives [14]. One possible approach is to exploit mappings from a reference scheme to search and navigate across a set of satellite vocabularies. For instance, Renardus and HILT [15] use the Dewey Decimal Classification (DDC). Some others prefer the Library of Congress Subject Headings (LCSH) [16, 17]. Both manual and semi-automatic solutions are proposed. Lauser et al. [18], with a focus on the agricultural domain, compare the two approaches and conclude that automatic procedures can be very effective but tend to fail when domain specific background knowledge is needed. Approaches to this problem have been proposed (see for instance [12]), but their accuracy still remains pretty low. It is therefore clear that automatic approaches typically require some form of manual validation [20], but limited work has been done in this direction and current interfaces to this purpose hardly scale with the size of the two ontologies [21, 22, 23]. A good survey of the state of the art in automatic tools for mapping computation can be found in [19], while the OAEI[5] initiative annually provides an evaluation of these tools.

For what said above, it is clear that when a very high accuracy is necessary purely manual or semi-automatic approaches, even if more time-consuming, are still prefera-

[1] www.fao.org/agrovoc/
[2] http://www.nlm.nih.gov/research/umls/
[3] http://en.wikipedia.org/
[4] http://www.bibliothek.uni-regensburg.de/projects/carmen12
[5] http://oaei.ontologymatching.org/

ble. Following this line, in this paper we present a technique for the semi-automatic mapping of generic categories to ontology concepts. As part of the proposed solution, we developed a tool - that we called SAM - that, as proved by our evaluation, allows saving around 50% of the time required by purely manual approaches.

The rest of this paper is organized as follows. Section 2 provides a motivating example showing the mapping process and typical problems that need to be faced. Section 3 describes the process of manual mapping. Section 4 presents the semi-automatic mapping approach and how the steps are supported by the SAM tool. Section 5 provides corresponding evaluation. Finally, Section 6 concludes the paper by summarizing the work done and outlining future work.

2 A motivating example

Consider the example in Fig. 1. It provides a small ontology where classes are represented with circles and individuals with squares; solid arrows represent relations between classes; dashed arrows represent relations between individuals or between an individual and corresponding class. Classes and relations between them constitute the TBox where the backbone is typically represented by *is-a* relations. Knowledge about the individuals forms the ABox where the relation between an individual and corresponding class is typically *instance-of*. Similarly to WordNet, each class can be associated a set of synonyms (here we do not provide definitions).

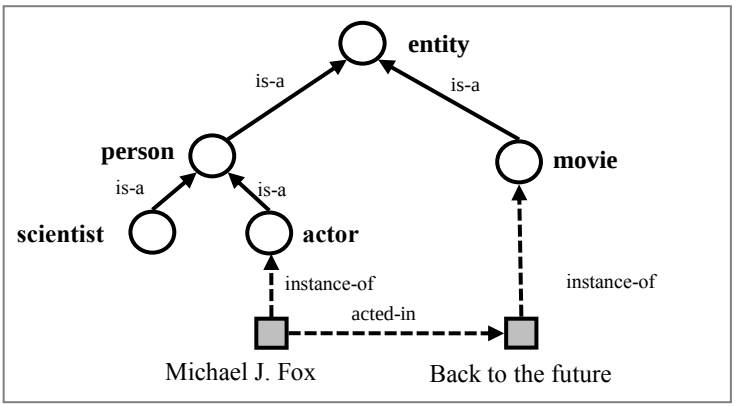

Fig. 1 – A sample target ontology

Suppose that our task is to extend the ontology by importing knowledge from the external source depicted in Fig. 2. As it often happens, the source is only partially structured, in the sense that none of its elements is explicitly marked as class, individual or relation. This makes the process of extracting knowledge approximate due to errors that might be made in interpreting them.

Existing knowledge extraction techniques, for instance those at the basis of YAGO, rely on the identification of known terms in the phrases denoting category names, i.e. terms that already appear as labels of concepts in the ontology we want to extend. This is done by first identifying what in linguistics is known as the *head of the phrase* and by mapping it with a concept in the ontology. For instance, the head of

Italian scientists is *scientist* (in its root form) which is in the ontology. This allows mapping them and enriching the ontology with the individuals extracted from the external source, thus importing *Antonio Meucci* as instance of *scientist*.

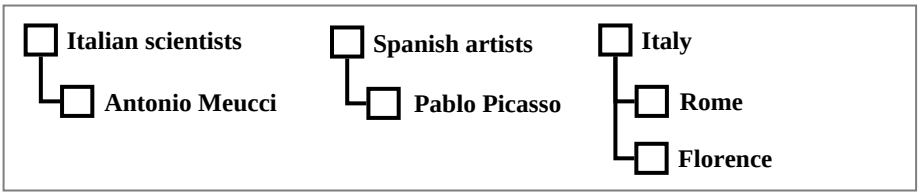

Fig. 2 – A sample external source

However, especially if automated, during this process many mistakes may arise. The identification of the head, which in turn is typically based on part of speech (POS) tagging, is an approximated process with accuracy that varies according to the tool and the dataset used to train it. For instance, POS tagging reaches 97.24% accuracy on the Penn Treebank WSJ dataset [1]. However, mistakes are amplified when the POS is used to identify the head. Even if the head is correctly identified, there might be cases in which the head is not in the ontology and cases in which more than one sense for it is available. In the former case, YAGO enriches the ontology by linking the category directly to the root of the ontology. For instance, since *artist* (the head of *Spanish artists*) is not in the ontology, *artist* is directly linked to *entity* (while a better choice would be *person*). In the latter case, YAGO as main heuristic selects the sense with higher rank in WordNet. Notice that the head of a phrase is always a common noun. The categories in which it is not present (for instance in the category *Italy* that is a proper noun) are simply ignored. The ontology that is obtained after the enrichment[6] is shown in Fig. 3.

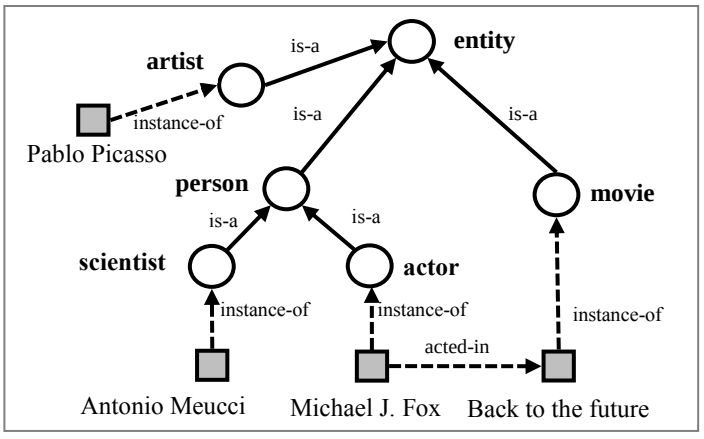

Fig. 3 – The enriched ontology

[6] Notice that here we imported only the head and the entities while in YAGO also the categories themselves are imported.

3 Manual mapping

The steps that we follow to manually map categories from an external source to the target ontology and consequently enrich it are as follows:

- **Step 1 - Filtering out proper nouns:** when the category represents a proper noun, it is marked as noise and filtered out. For example the category *Crittenden County* represents the name of a place in Arkansas, USA.
- **Step 2 - Identification of the head of the category:** when the category is a single word, it is clearly selected as head. When the category is constituted by more than one word the head is manually selected. For instance, for the category *Iraqi Sunni Muslims* the word *Muslim* is selected, while for the category *racehorses trained in Italy* the word *racehorse* is selected. In some cases a multiword has to be selected. For instance, for the category *amusement parks in New England* the multiword *amusement park* is selected.
- **Step 3 - Mapping the head to a concept:** a suitable concept corresponding to the head is searched in the target ontology. Possible candidates are evaluated and one of them is selected trying to understand the most suitable one. If a good one is found the process is completed.
- **Step 4 - Creation of a new concept:** if no suitable concept is found for the head of the category, an additional external vocabulary is used to determine a definition for it. A good definition should provide *genus et differentia*, i.e. it should provide information about the kind (the genus) and how it differs from the kind (the differentia). For example, *pathologist* can be defined as *a scientist who studies parasites and their biology and pathology* where scientist is the genus and the rest represents the differentia.
- **Step 5 - Identification of the parent:** similarly to Step 3, the genus in the definition is used to identify a suitable parent concept, if any, in the target ontology. For example, *scientist* is a suitable parent for *pathologist*.
- **Step 6 – Enriching the target ontology:** by using the mapping between the categories and the concepts in the ontology (either as direct mapping or through a parent) the ontology is enriched with new concepts and corresponding individuals extracted from the external source.

To evaluate the potential of this method, we took WordNet as target ontology and YAGO as external source. In fact, even if YAGO is already the result of a mapping between Wikipedia and WordNet, we found out that 15,480 Wikipedia categories were directly mapped to the root concept *entity* of WordNet. We applied the steps above to 2,000 of these categories randomly selected. The results are provided in Table 1 and show that:

- 18% of the mapped categories are actually proper nouns
- 56% of the categories can be mapped to a more specific concept in WordNet
- 26% of the categories can be better mapped if a new concept is created and mapped to an existing parent concept in WordNet

Categories analyzed	Proper Nouns	Concepts found	Concepts created
2000	358	1120	522

Table 1. Results of the manual mapping

4 Semi-automatic Mapping

By applying the manual steps we can clearly achieve a very accurate mapping and enrichment, however at the price of a higher cost in terms of human resources and time needed. To overcome this limitation, we developed the SAM tool (implemented in Java) to assist the user (typically an ontology expert) and partially automate the necessary steps. The steps remain pretty much the same but the process is preceded by a preprocessing phase during which the system is trained in order to automatically recognize a category as proper noun or in alternative to identify its head. For each of the categories, the steps are as follows:

- **Step 1 - Filtering out proper nouns:** if the system recognizes the category as a proper noun no head is computed. The user is free to accept the suggestion or proceed to the next step.
- **Step 2 - Identification of the head of the category:** the system computes a head for the category. The user is free to accept the suggestion or provide an alternative one.
- **Step 3 - Mapping the head to a concept:** since the system keeps track of previous choices made by the user, if the head of the category corresponds to a word which has been already processed in the past then the system suggests previously assigned concepts. To help the user deciding, it shows them in a list with corresponding categories. For instance, in processing the category *Mexican Americans* and by automatically identifying *American* as head, it pops up the information that this head appeared in the previously processed category *Jamaican Americans in the Unites States Military* such that the same concept can be selected. If no similar cases are found or none of them is considered relevant by the user, the system looks up in the target ontology to identify the concepts corresponding to the head. They are given in a list as shown in Fig. 4. The user can pick one of them or, if none is found or none of them is considered correct, move to the next step.
- **Step 4 - Creation of a new concept:** the system queries an external vocabulary to identify useful information that the user can utilize to determine a suitable definition for the head.
- **Step 5 - Identification of the parent:** similarly to Step 3, the genus of the definition provided by the user is used to look into the target ontology for candidate concepts for the parent. The user is free to select one of the suggestions or reject them. If none of them is considered appropriate, the system asks for an alternative definition by coming back to step 4. A new concept is otherwise generated by the system and linked to the corresponding parent in the target ontology.

- **Step 6 – Enriching the target ontology:** At the end of the process, the entities associated to the category are automatically used to populate the ontology.

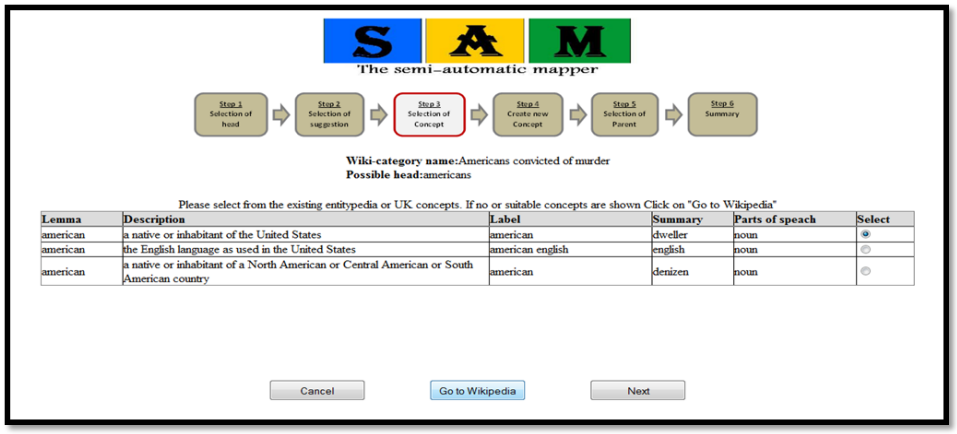

Fig. 4 – A snapshot of the SAM interface

5 Evaluation

To evaluate SAM we used Entitypedia [10] as target ontology and the 15,480 categories of YAGO that were directly mapped to *entity* as external resource. Wikipedia was used as external vocabulary at Step 4. Developed at the University of Trento in Italy, Entitypedia is a knowledge base with a precise split between individuals (the ABox), classes, attributes and relations (the TBox) and their lexicalization as proper nouns and common nouns, respectively. Entitypedia is progressively extended by collecting knowledge from several sources, including WordNet.

With the pre-processing, the 15,480 categories of YAGO were POS tagged by using the Stanford NLP POS tagger [1]. After the tagging, a set of patterns were identified in order to automatically recognize the head. This was done by looking at common noun plural tags (/NNS). 8,998 categories were found to have exactly one such tag; 292 of them with more than one; 6,190 of them do not have any. In the first case the corresponding word was selected as head; in the second and third case 7 and 33 different patterns were identified respectively and for each of them a different choice was made also taking into account the mistakes made by the POS tagger. Patterns are also used to identify proper nouns. An example of pattern is:

{JJ}+ {NNP}* {NNPS}+

where {JJ}+ indicates one or more adjectives, {NNP}* zero or more proper nouns and {NNPS}+ one or more common nouns in plural form. An example of category matching this pattern is *Indian Zen Buddhists* with *Buddhist* its head. By evaluating a sample of 500 categories we found that this approach leads to an accuracy of 98.4% (only 8 mistakes).

The evaluation process comprises of 2 parts. During the first part, a trained user was given 200 YAGO categories randomly selected to be mapped manually. During the second part, the user was asked to use SAM to map 200 new YAGO categories different from the previous ones. By trained user we mean a user who was familiar with the manual mapping process as he was involved in the analysis phase described in Section 3 but not at all familiar with SAM. In other words, the user neither participated to the design nor to the implementation of the tool. The user was given precise evaluation guidelines including clear steps about the tasks to be done and was monitored during the whole experiment. Table 2 provides some examples of mapped categories. In the table, the head of each category is given in bold followed by either the concept found in the target ontology or the definition of the new concept otherwise.

Category: Futurologists *Concept found*: - *Concept created*: futurologist (scientist and social scientist whose speciality is to attempt to systematically predict the future, whether that of human society in particular or of life on earth in general)
Category: Rare **diseases** *Concept found*: disease (an impairment of health or a condition of abnormal functioning) *Concept created*: -
Category: Landforms of Turkey *Concept found*: - *Concept created*: landform (is largely defined by its surface form and location in the landscape)
Category: Germans of Polish descent *Concept found*: german (a person of German nationality) Concept created: -
Category: Pharaohs of the Twenty-sixth dynasty of Egypt *Concept found*: pharaoh (the title of the ancient Egyptian kings) *Concept created*: -
Category: Roman Catholic **dioceses** in the Holy Roman Empire *Concept found*: diocese (the territorial jurisdiction of a bishop) *Concept created*: -
Category: Recipients of the Distinguished Service Cross (United States) *Concept found*: recipient (a person who receives something) *Concept created*: -
Category: Sexually transmitted **diseases** and infections *Concept found*: disease (an impairment of health or a condition of abnormal functioning) *Concept created*: -

Table 2. Examples of mapped categories

With the evaluation, we took note of the number of proper nouns, concepts found and new concepts created as well as of the time taken. Figures for both the first and second part of the experiment are reported in Table 3.

Mapping	Manual Mapping	Semi-automatic Mapping
Proper Nouns	11	12
Concepts found	122	111
Concepts created	67	77
Amount of Time (minutes)	169.22	89.98

Table 3. Manual and semi-automatic mapping compared

As it can be noticed from the table, the distribution of the different cases is slightly different. For instance, during the semi-automatic mapping more concepts had to be created. As it can be noted from the description of the steps, these cases are those requiring more time. Nevertheless, with the help of the tool the user was able to complete the process in around half of the time.

6 Conclusions

In this paper we have shown that the process of manually enriching ontologies with knowledge coming from external sources can be significantly speed up, still guaranteeing a high level of accuracy, by using tools that interactively support the user during the mapping phase. In fact, our experiments show that by using the SAM tool it is possible to save around 50% of the time needed by purely manual approaches.

As future work we plan to conduct accurate usability studies on the user interface of SAM to identify critical parts that can be improved to facilitate or further speed up the process. The patterns used to identify proper nouns and the head of the categories will be tested against a broader set of categories to verify how the accuracy varies on unseen data and to eventually extend the number of patterns. SAM has been customized to work on YAGO (input) and Entitypedia (output), while a future extension may allow generalizing the input/output.

Acknowledgment

The research leading to these results has received funding from the CUbRIK Collaborative Project, partially funded by the European Commission's 7th Framework ICT Programme for Research and Technological Development under the Grant agreement no. 287704. We would like to thank Professor Fausto Giunchiglia for his constant guidance and Suresh Daggumati for the evaluation.

References

1. K. Toutanova, D. Klein, C. Manning, Y. Singer, 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. HLT-NAACL, pp. 252-259.
2. F. Giunchiglia, D. Soergel, V. Maltese, A. Bertacco, “Mapping large-scale Knowledge Organization Systems” 2nd International Conference on the Semantic Web and Digital Libraries (ICSD), 2009.

3. R. Studer, V. R. Benjamins, D. Fensel, “Knowledge engineering: principles and methods”, Data and Knowledge Engineering, 25, 161–197, 1998.
4. T. R. Gruber, “A translation approach to portable ontology specifications”, Knowledge Aquisition, 5 (2), pp. 199–220, 1993.
5. F. Baader, D. Calvanese, D. McGuinness, D. Nardi, P. F. Patel-Schneider, “The Description Logic Handbook: Theory, Implementation and Applications”, Cambridge University Press, 2002.
6. C. Fellbaum, “WordNet: An Electronic Lexical Database”, MIT Press, 1998.
7. C. Matuszek, J. Cabral, M. Witbrock, J. DeOliveira, “An introduction to the syntax and content of Cyc”, AAAI Spring Symposium, 2006.
8. A. Pease, G. Sutcliffe, N. Siegel, S. Trac, “Large theory reasoning with SUMO at CASC”, AI Communications, 23(2-3), pp. 137–144, 2010.
9. F. M. Suchanek, G. Kasneci, G. Weikum, “YAGO: A Large Ontology from Wikipedia and WordNet”. Journal of Web Semantics, 2011.
10. F. Giunchiglia, V. Maltese, B. Dutta, “Domains and context: first steps towards managing diversity in knowledge”. Journal of Web Semantics, special issue on Reasoning with Context in the Semantic Web, 2012. DOI: 10.1016/j.websem.2011.11.007.
11. N. Noy, “Semantic Integration: A survey of ontology-based approaches”. SIGMOD Record, 33(4), pp. 65–70, 2004.
12. F. Giunchiglia, P. Shvaiko, M. Yatskevich, “Discovering missing background knowledge in ontology matching”. Proceedings of the 17th European Conference on Artificial Intelligence (ECAI), pp. 382–386, 2006.
13. T. Koch, H. Neuroth, M. Day, 2003. “Renardus: Cross-browsing European subject gateways via a common classification system (DDC)”. In I.C. McIlwaine (Ed.), Subject retrieval in a networked environment. Proceedings of the IFLA satellite meeting held in Dublin - IFLA Information Technology Section and OCLC, pp. 25–33.
14. D. Vizine-Goetz, C. Hickey, A. Houghton, R. Thompson, 2004. “Vocabulary Mapping for Terminology Services”. Journal of Digital Information 4(4)(2004), Article No. 272.
15. D. Nicholson, A. Dawson, A. Shiri, 2006. “HILT: A pilot terminology mapping service with a DDC spine”. Cataloging & Classification Quarterly, 42 (3/4). pp. 187-200.
16. C. Whitehead, 1990. “Mapping LCSH into Thesauri: the AAT Model”. In Beyond the Book: Extending MARC for Subject Access, pp. 81.
17. E. O'Neill, L. Chan, 2003. “FAST (Faceted Application for Subject Technology): A Simplified LCSH-based Vocabulary”. World Library and Information Congress: 69th IFLA General Conference and Council, 1-9 August, Berlin.
18. B. Lauser, G. Johannsen, C. Caracciolo, J. Keizer, W. R. van Hage, P. Mayr, 2008. “Comparing human and automatic thesaurus mapping approaches in the agricultural domain”. Proc. Int’l Conf. on Dublin Core and Metadata Applications.
19. J. Euzenat, C. Meilicke, H. Stuckenschmidt, P. Shvaiko, C. Trojahn, 2011. “Ontology Alignment Evaluation Initiative: six years of experience”. Journal on Data Semantics.
20. V. Maltese, F. Giunchiglia, A. Autayeu, 2010. “Save up to 99% of your time in mapping validation”. 9th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE).
21. G. G. Robertson, M. P. Czerwinski, J. E. Churchill, 2005. “Visualization of mappings between schemas”. SIGCHI Conference on Human Factors in Computing Systems.
22. A. Halevy, 2005. “Why your data won’t mix”. ACM Queue, 3(8), 50–58.
23. S. Falconer, M. Storey, 2007. “A cognitive support framework for ontology mapping”. International Semantic Web Conference (ISWC).