Data-driven Schema Matching in Agricultural Learning Object Repositories

Antonis Koukourikos[1,4], Giannis Stoitsis[2,3], Pythagoras Karampiperis[4]

[1]University of Piraeus, Department of Digital Systems. 80, Karaoli and Dimitriou Str, Piraeus, Greece, 18534
[2]Agro-Know Technologies, Grammou 17, Vrilissia, Athens, Greece, 15235
[3]Universidad de Alcalá, Pza. San Diego, s/n - 28801 Alcalá de Henares, Madrid, Spain
[4]Software and Knowledge Engineering Laboratory, Institute of Informatics and Telecommunications, National Center for Scientific Research "Demokritos"Agia Paraskevi Attikis, P.O.Box 60228, 15310 Athens, Greece

Abstract. As the wealth of structured repositories of educational content for agricultural object is increasing, the problem of heterogeneity between them on a semantic level is becoming more prominent. Ontology matching is a technique that helps to identify the correspondences on the description schemas of different sources and provide the basis for interesting applications that exploit the information in a linked fashion. The present paper presents a data-driven approach for discovering matches between different classification schemas. The approach is based on content analysis and linguistic processing in order to extract information in the form of relation tuples, use the extracted information to associate the content of different repositories and match their underlying classification schemas based on the degree of content similarity. The preliminary results verified the validity of the approach, as both experiments produced a semantically valid matching in 68% of the examined classes. The results also exposed the need for refinements on the linguistic processing of the available textual information and on the definition of relation similarity, as well as, the need to exploit structural information in order to move from discovering semantically valid matches to effectively handling class specializations and generalizations.

Keywords: ontology matching, classification schemas, educational content, agricultural objects

1 Introduction

The progress on the availability and structuring of online information has made available huge amounts of disjoint information for multiple domains. The usability and effectiveness of this information is greatly increased if the contributions of different content providers is associated and used in liaison with each other. Therefore, the problem of managing the heterogeneity between various information resources in order to integrate seamlessly and efficiently the underlying knowledge is of particular interest.

In the context of the Semantic Web, ontologies are a common medium for describing the domain of interest and providing a contextualization of the different terms used for specifying the characteristics of the involved entities. Ontology matching is one of the prominent technologies used for integrating such descriptions on the conceptual level However, information and knowledge resources are not always associated with an ontology. Classifications of different complexity and formalization are employed in different repositories. Some of them do deploy full-fledged ontologies, usually expressed in OWL or RDFS, while others use less complex solutions, like XML schemas or simple categorization.

The present paper builds on the ideas from the fields of ontology matching and discusses a data-driven approach towards the consolidation of the schemas describing different repositories. Our approach is based on the notion that documents from different repositories that discuss similar subjects are likely to have corresponding classifications in their respective schemas. Therefore, by extracting and comparing relations from these documents we are able to identify alignments between the schemas. Furthermore, additional restrictions can be posed after taking into account the hierarchical structure of the compared classifications. We formalized the above hypothesis and tested its application using repositories of educational content for agricultural objects.

The rest of the paper is structured as follows: We present some of the popular techniques of ontology alignment and relevant systems. Afterwards, a description of our method and the data collections that we used is provided. Next, the obtained results for the examined datasets are presented. We conclude with an indication of rooms for improvement and report future steps for calibrating and expanding on the existing infrastructure.

2 Related Work

The purpose of ontology matching is, in a broad context, to define correspondences and mappings between concepts, as the latter are expressed in different conceptualization schemas. Several formalizations of the above statement have been proposed [1, 2, 3]. In [4] it is stated:

Let O1 and O2 distinct ontologies. An alignment between these ontologies is a set of correspondences between entities belonging to the two ontologies. A correspondence is a quadruple of the form:<id,e1,e2,r>, where:

- id is a unique identifier for the correspondence
- e1 is an entity of the first ontology O1
- e2 is an entity of the second ontology O2
- r is the type of relation between e1 and e2

The relation between the matched entities can be equivalence, generalisation/specialisation and others, depending on the nature of the problem that is being examined.

There are various techniques used for performing ontology matching. A common method is the application of linguistic analysis within the ontology in order to compute similarities on the textual level. Another strategy for ontology matching is the examination of structural properties of the ontologies to be merged. The graph structure derived from the ontology, commonly via is-a/ part-of relationships between concepts, provides a means for examining the similarity between two ontologies based on the connections between their concepts. Instance-based approaches, where the objects described by the ontologies are available and annotated with ontological terms, are also of particular interest. Similarity between instances can lead to suggestion of similarities between the underlying concepts. Finally, external knowledge information, such as thesauri, dictionaries and taxonomies, are frequently employed in ontology matching in order to provide further information about the semantics of the concepts and relations in the ontologies to be matched.

In practice, these approaches are not mutually exclusive, as ontology alignment systems can use combinations of them or employ selection strategies to invoke a matcher based on features specific to the matching task at hand. Some prominent recent alignment systems and their approaches are described below.

SAMBO [5] is used for matching (and merging) biomedical ontologies. It supports the merging of ontologies expressed in OWL format. The system combines different matchers, each one computing a similarity value in the [0, 1] space. The terminological matcher examines similarities between the textual descriptions of concepts and restrictions of the ontologies, using the n-gram and edit distance metrics and a linguistic algorithm that compares the lists of words of which the descriptions terms are composed and discovers the common words. A structural matcher relies on the position of concepts relative to already aligned concepts and iteratively aligns additional entities based on their structural association (is-a/part-of connections with entities aligned during a previous iteration). SAMBO also examines the similarity of terms in the ontologies with an external domain-specific resource (UMLS) and employs a learning matcher that classifies documents with respect to their relation with ontology concepts and associates the entities that encapsulated the same documents.

RiMOM [6] uses a multi-strategy ontology matching approach. The matching methods that are employed are (a) linguistic similarity and (b) structural similarity. The linguistic similarity adopts the edit distance and vector instance metrics, while the structural similarity is examined by a modified similarity flooding [7] implementation. For each matching task, RiMOM quantifies the similarity characteristics between the examined ontologies and dynamically selects the suitable strategy for performing the task.

The ASMOV [8] system handles pairs of ontologies expressed in OWL. The process employed by ASMOV includes two distinct phases. The similarity calculation phase activates linguistic, structural and extensional matchers in order to iteratively compute similarity measures for each pair of entities comprised by the elements of the ontologies to be matched. The measures are then aggregated into a single, weighted average value. From this phase, a preliminary alignment is produced by selecting the maximum similarity value for each entity. During the semantic verification phase, this

alignment is iteratively refined via the elimination of the correspondences that are not verified by assertions in the ontologies.

BLOOMS [9] is an alignment system that discovers schema-level links between Linked Open Data datasets by bootstrapping already present information from the LOD cloud. After a light-weight linguistic processing, it feeds the textual descriptions of concepts in two ontologies to the Wikipedia search Web Service. The Wikipedia categories to which the search results belong to are inserted into a tree structure that is expanded with the subcategories of the aforementioned categories, until the tree reaches the fourth level. The trees belonging to the "forests" of the two input ontologies are compared in pairs and an overlap value is assigned to each tree pair. Based on this value, BLOOMS defines equivalence and specialization relations between the concepts of the ontologies.

The aforementioned systems have produced significant results in the context of classification schema matching. However, they mostly handle schemas expressed in a specific format (e.g. OWL ontologies), so they require a certain level of conformance in order to perform the matching task. Repositories of learning content for agriculture, however, use a wide variety of different formalizations for classifying their content and their metadata. The amount of different approaches [10, 11] and the variability on the methodologies and lexicalization [12, 13] pose several interesting issues for the efforts of ensuring that the crucial need for interoperability is met. Our approach aims to exploit the strategies employed in ontology alignment in order to develop a system that is able to match classification schemas expressed in different ways. In order to remedy the inability of direct comparison of the schemas due to their different formats, we consider the examination of the underlying actual data as a means for discovering the semantic associations behind the different classifications.

3 Methodology

Our approach focuses on the analysis of the actual educational objects described by the classification schemas to be merged. The specific goal of our experiments was to match each of two distinct description schemas with a third one, that is, to perform two independent, one-to-one matching tasks. The base schema for our experiments was the one use by the Organic.Edunet Web portal. In the first run, we applied our method for the base schema and the schema of OER Commons Green. For the second run, we matched the Organic.Edunet ontology and the taxonomy of Organic Eprints.
The following subsections describe (a) The datasets that we used and their description schemas, (b) the process of extracting information from the datasets and (c) the execution of the matching task.

3.1 Datasets

The schema that was used in both runs of our experiment was the ontology of the Organic.Edunet Web portal (http://www.organic-edunet.eu). The Organic.Edunet Web portal for agricultural and sustainable education was launched in 2010. Its aim

has been to facilitate access, usage and exploitation of digital educational content related to Organic Agriculture (OA) and Agroecology (AE). In order to achieve this aim, it networked existing collections with educational content on relevant topics from various content providers, into a large federation where content resources are described according to standard-complying metadata. The underlying description schema, the Organic.Edunet organic ontology, is expressed in OWL. An example of a metadata record from the Organic.Edunet repository is depicted in Figure 1.

```
<lom xmlns="http://ltsc.ieee.org/xsd/LOM">
<general>
  <identifier></identifier>
  <title><string language="en">Insulating livestock and other farm buildings</string></title>
</general>
<technical>
  <format>text/html</format>
  <location>http://www.ces.purdue.edu/extmedia/AE/AE-95.html</location>
</technical>
<taxonPath>
   <source>
     <string language="en">Organic.Edunet Ontology</string>
   </source>
   <taxon>
     <id>http://www.cc.uah.es/ie/ont/OE-Predicates#ProvidesNewInformationOn :: http://www.cc.uah.es/ie/ont/OE-OAAE#LivestockHousing</id>
     <entry>
       <string>ProvidesNewInformationOn :: LivestockHousing</string>
     </entry>
   </taxon>
</taxonPath>
</lom>
```


Fig. 1. Snippet of a metadata entry from Organic.Edunet.

On the first run of the system, the second input classification schema was the one of the OER Commons Green repository (http://www.oercommons.org/green). OER Commons Green is part of OER Commons, which was created by ISKME (http://www.iskme.org/) as a way to provide support for and build a knowledge base around the use and reuse of open educational resources. From the content available via OER Commons Green, there are currently 3157 documents organized by subject in a 2-tier hierarchy.

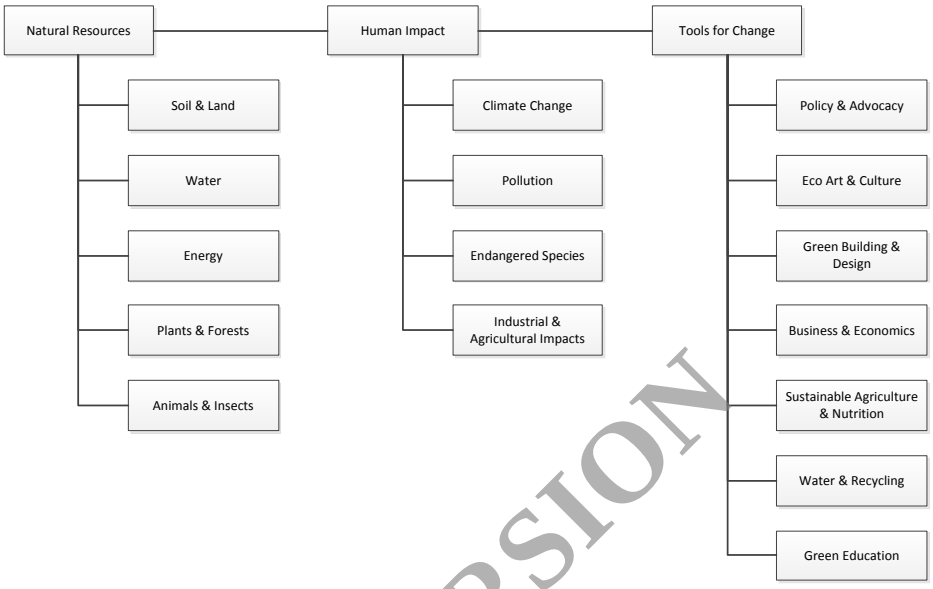

Fig. 2. The Green OER classification of educational content by subject.

The XML metadata descriptions of the educational objects included in the Green OER Commons repository use external classification methods to provide additional taxonomical information for each object. Since our goal is to match the central taxonomy of Green OER Commons to that of the Organic.Edunet portal, we did not take into account these associations. Instead, we produced an augmented version of the original XML file with an added <goer_subject> element for each document, based on its classification at the Green OER Commons site. For adding the new element, we parsed the web page dedicatied to the document in order to extract the classification lexicals. Then, we manipulated the DOM of the XML file using the standard Java libraries for the task, inserted a new node for the element with the Green OER classification as its value and serialized the final DOM into the final XML document.

The second experiment aimed to match the classification of Organic.Edunet with that of Organic Eprints (http://www.orgprints.org/). The Organic Eprints archive has been developed by the International Centre for Research in Organic Food Systems (http://www.icrofs.org/) and its main goal is to facilitate the communication and dissemination of research objects in the domain of organic agriculture. Its classification system is based on subject definition for the included material. The maximum depth of the classification is 4; however, most categories reach level 2 at most. The subjects of each object are declared in its metadata description via one or more dc:subject elements.

```
<record>
<header>
<identifier>oai:orgprints.org:122</identifier>
</header>
<metadata>
<dc:subject> Crop combinations and interactions</dc:subject>
<dc:subject> Cereals, pulses and oilseeds</dc:subject>

</metadata>
</record>
```

Fig. 3. Snippet of an object description in the Organic Eprints record list.

3.2 Information Extraction

The first step on the implementation of our system is to extract a relation set from the documents within the datasets. At this stage, we take into account educational content solely in the English language and we handle documents in the DOC, PDF and HTML formats. Using the metadata descriptions for the records in each repository, we selected 500 entries from each one. The XML file for each record set was parsed and the record entries were stored as objects in a linked list. We selected random indexes from the lists and, as long as the format and language of the corresponding document were suitable and we had not reached the amount limit, we retrieved the actual resource for further processing.

Before performing the information extraction task some pre-processing of the ontological elements and the examined documents was necessary. We applied some lightweight linguistic processing to the textual descriptions of the classification schemas, in order to obtain proper terms and capitalization. For the documents retrieved in HTML format, we applied a boilerplate removal module in order to exclude formative content (tags, scripting snippets) and content irrelevant to the interesting content (menus, advertisements, comments etc.). The module was built on top of the boilerpipe library (http://code.google.com/p/boilerpipe). A step that was deemed necessary was the resolution of co-references within the text. The absence of such an analysis led to the production of numerous relations that were not useful since they associated entities that could not be resolved. Use of pronouns and generic terms, like "the band", "the group" do not allow the direct expansion of the relation set for an entity. To overcome this issue, we use the co-reference resolution module of the OpenNLP Tools (http://opennlp.apache.org/).

For the information extraction process, we used the REVERB system [14], which follows the Open Information Extraction paradigm, building on the methodology of previous systems, like TEXTRUNNER [15]. TEXTRUNNER returns a set of relation tuples by executing a single pass over the entire input corpus and assigns a probability to each tuple based on the probabilistic model of redundancy in text proposed by [16]. REVERB expands this method and introduces a constraint enforcement mechanism in order to improve on the accuracy of the produced relation set. A syntactic constraint

eliminates incoherent and uninformative extractions, while a lexical constraint rejects overly specific - and thus not useful - relations by examining the amount of distinct arguments presented in the corpus for the relation. The relation tuples produced by REVERB include three components. That is, they have the form (Arg1, Rel, Arg2), where Arg1 is an entity connected unidirectionally via the relation Rel with Arg2.

The use of an open information extraction module, as opposed to a domain-aware system, can be somewhat detrimental to the precision of the results set; however, it constitutes the system adaptable to radically different content and allows its usage for educational objects of different domains.

We ran the information extraction system for each document in the three collections. The results are presented in Table 1.

Dataset	Retrieved Relations
Organic.Edunet	28,620
OER Commons Green	21,462
Organic Eprints	30,583

Table 1. Amount of relations obtained from each document collection.

3.3 Alignment

We define the following cases for pairs of relation triples that are likely to have some semantic relevance between them:

- Relations triples with similar Rel fields but different Arg fields
- Relation triples with similar Arg fields but different Rel fields
- Relation triples with all three fields similar

At this version of the system, we use a pure linguistic approach for deciding on the similarity between the fields of a triple. First, the corresponding fields are stemmed and possible secondary terms (prepositions, auxiliary verbs etc.) are eliminated. Then, we retrieve the senses to which the resulted terms belong according to WordNet (http://wordnet.princeton.edu/). If the terms under comparison shared a common WordNet sense, we consider the terms similar. The notion of "sense" includes the case of synonyms and to some extend covers related terms, though domain-specific relations are not always discovered.

The next step is the calculation of a similarity score for each pair of documents from the educational repositories. The relations retrieved from each document are compared and a similarity score is generated for each document pair. The score is calculated following the formula

$$RelSim(d1, d2) = \sum_{i=1}^{N} \sum_{j=1}^{M} Comp(t_i, t_j)$$

where d1 and d2 are the compared documents and ti, tj are relation triples retrieved from d1 and d2 respectively. The function for determining the similarity between triples was straight-forward for this implementation of our method. We assigned a

value of 1 when all fields were similar, a value of 0.66 when the argument fields were similar and a value of 0.33 when only the relation field was similar between the triples.

The aforementioned process produces a similarity score matrix for the examined documents. The document similarity scores are then transferred to the classification similarity by associating the documents with their already existing categorizations. Let C1 be a class of the first classification schema and C2 a class of the second classification schema. The system (a) locates the documents that are classified under C1 and those that are classified under C2, (b) retrieves the similarity scores of each pair constructed from these subsets of documents and (c) adds their similarity scores. Formally,

$$ClassSim(C1, C2) = \sum_{i=1}^{N} \sum_{j=1}^{M} RelSim(d_i, d_j)$$

where di is classified under C1 and dj is classified under C2. A bigger ClassSim score indicates a higher probability that C2 is a classification equivalent to C1.

4 Results

In this section we will present the results of the described system for two distinct matching tasks. The first run aimed to match the classification schema of OER Commons Green with that of Organic.Edunet. The second execution matched the classification schema of Organic.Eprints against the Organic.Edunet schema. For each task, we provide the amount of similar relation tuples between the two document collections employed, the average RelSim scores produced. The results were compared to the manual matching between the used classification schemas.

4.1 Matching between Organic.Edunet and OER Commons Green

As mentioned, the REVERB system returned 28,620 relations for the Organic.Edunet document collection and 21,462 for the OER Commons Green document collection. The similarities on the relation tuple level are summarized in Table 2.

Relation Similarity Case	Total Amount	Average	Average RelSim
All fields similar	768	1.536	9.285
Argument fields similar	2,167	4.334	
Relation fields similar	7,407	14.814	

Table 2. Related relation tuples between the Organic.Edunet and OER Commons Green document collections

From the 19 subjects included in the OER Commons Green classification, 6 were deemed equivalent to the appropriate Organic.Edunet class. 7 of the subjects were associated with a class that should be its generalization as opposed to its equivalent.

Finally, 6 of the OER Commons Green subjects were not matched to the most suitable Organic.Edunet class. In the latter case, the ClassSim score for the most appropriate was the third larger in the two worst cases and the second larger in the remaining four cases. Overall, the matching was semantically correct in 68.4% of the examined classes.

4.2 Matching between Organic.Edunet and Organic Eprints

The comparison of the 28,620 relations derived from Organic.Edunet and the 30,583 relations derived for Organic.Eprints produced the results presented in Table 3.

Relation Similarity Case	Total Amount	Average	Average RelSim
All fields similar	814	1.628	9.042
Argument fields similar	2,416	4.832	
Relation fields similar	6,402	12.804	

Table 3. Related relation tuples between the Organic.Edunet and Organic Eprints document collections

The Organic Eprints classification system includes 66 topics and subtopics. 13 of the subjects were associated with the appropriate Organic.Edunet class by the matching system. 32 were deemed equivalent to an Organic.Edunet class that is semantically their specialization or generalization, while 21 subjects were matched erroneously. Similarly to the Green OER Commons case, 68.1% of the classes were matched correctly with respect to their semantics.

5 Conclusions and Future Work

The preliminary results of our experiment indicate that there are adequate reasons to further investigate the described approach. However, there is significant work that needs to be done in order to improve on the system. The calculation of relation triple similarity seems to be the most detrimental aspect for the accuracy of the system. The simple textual analysis that was employed for determining similar fields in the triples does not produce a reliable response in all cases, as it is both imprecise and incomplete. Specifically, it seems that a common relation field is not a good indicator of a semantic association per se, so its contribution to the similarity score should be reduced. We will examine the use of external information, like the AgroVoc vocabulary (http://aims.fao.org/website/AGROVOC-Thesaurus/sub) for the specific use case, and more elaborate techniques like entity matching on the argument fields of the relations in order to improve the efficiency of this step. Furthermore, we will observe the impact of introducing domain knowledge on the precision of the information extraction process. The need to retain the open, domain-independent nature of the information extraction process is important to us; however, approaches like the ones proposed by [17] and [18] allow the unobtrusive inclusion of domain-specific information.

An important issue that has not been addressed adequately by the current method is distinguishing between equivalence and specialization relations between the compared classes. An important step is to combine our metrics with the existing structural information in the classification schemas so as to identify the exact type of association between two classes.

An interesting outcome from the examination of our results is that there is a significant amount of relations that deal with objects relevant to the domain but do not seem to be covered by the existing classifications. In the future, we will examine in more detail the abilities of our approach in terms of automatically enriching and populating the classification schemas of the repositories, leading to a more accurate and fine-grained description for the specific domain.

6 Acknowledgments

The research leading to these results has received funding from the European Union Seventh Frame-work Programme, in the context of the SemaGrow (ICT-318497) project.

This paper also includes research results from work that has been funded with support of the European Commission, and more specifically the project CIP-ICT-PSP-270999 "Organic.Lingua: Demonstrating the potential of a multilingual Web portal for Sustainable Agricultural & Environmental Education" of the ICT Policy Support Programme (ICT PSP).

7 References

1. Kalfoglou, Y. & Schorlemmer, M. (2003). Ontology mapping: the state of the art. The Knowledge Engineering Review, 18(1), pp. 1-31.
2. Shvaiko, P. & Euzenat, J. (2005). A survey of schema-based matching approaches. Journal on Data Semantics, IV, pp. 146-171.
3. Zimmermann, A., Krötzsch, M., Euzenat, J. & Hitzler, P. (2006). Formalizing ontology alignment and its operations with category theory. In proceedings of the 4th International Conference on Formal Ontology in Information Systems (FOIS), pp. 277-288.
4. Euzenat, J & Shvaiko, P. (2007). Ontology Matching. Springer.
5. Lambrix, P & Tan, H. (2006). SAMBO – a system for aligning and merging biomedical ontologies. Journal of Web Semantics, 49(1), pp. 196-206.
6. Li, J., Tang, J., Li, Y., & Luo, Q. (2009). Rimom: A dynamic multistrategy ontology alignment framework. IEEE Transactions on Knowledge and Data Engineering, 21(8), pp. 1218–1232.
7. Melnik, S., Garcia-Molina, H. & Rahm, E. (2002). Similarity flooding: a versatile graph matching algorithm. In proceedings of the 18th International Conference on Data Engineering (ICDE), pp. 117-128.
8. Jean-Mary, Y.R., Shironoshita, E.P. & Kabuka, M. R. (2009). Ontology matching with semantic verification. Journal of Web Semantics, 7(3), pp. 235-251.
9. Jain, P., Hitzler, P., Sheth, A.P., Verma, K., & Yeh, P.Z. (2010). Ontology Alignment for Linked Open Data. In: proceedings of the 9th Intl Semantic Web Conference (ISWC2010).

10. Palavitsinis N. & Manouselis, N. (2009). A Survey of Knowledge Organization Systems in Environmental Sciences. I.N. Athanasiadis, P.A. Mitkas, A.E. Rizzoli & J. Marx-Gómez (eds.), Information Technologies in Environmental Engineering, Proceedings of the 4th International ICSC Symposium, Springer Berlin Heidelberg, 2009.
11. Palavitsinis, N. & Manouselis, N. Agricultural Knowledge Organisation Systems: An Analysis of an Indicative Sample. Sicilia M.-A. (Ed.), Handbook of Metadata, Semantics and Ontologies, World Scientific Publishing Co. (in press).
12. Manouselis, N., Najjar, J., Kastrantas, K., Salokhe, G., Stracke, C.M., & Duval, E., (2010), Metadata interoperability in agricultural learning repositories: An analysis, Computers and Electronics in Agriculture, 70(2), March, pp. 302-320
13. Manolis, N., Kastrantas, K. & Manouselis, (2012) N. Revisiting an analysis of agricultural learning repository metadata: preliminary results. Proceedings of the 6th Metadata and Semantics Research Conference (MTSR'12), Cádiz, Spain, 28-30 November 2012.
14. Etzioni, O., Fader, A., Christensen, J., Soderland, S., & Mausam (2011). Open Information Extraction: the Second Generation. International Joint Conference on Artificial Intelligence.
15. Yates, A., Cafarella, M., Banko, M., Etzioni, O., Broadhead, M., & Soderland, S. (2007). TextRunner: Open Information Extraction on the Web. Computational Linguistics, 42.
16. Downey, D., Etzioni, O., & Soderland, S. (2005). A probabilistic model of redundancy in information extraction. In proceedings of International Joint Conferences on Artificial Intelligence (IJCAI-05), pp. 1034-1041.
17. Soderland S., Roof B., Qin B., Xu S., Mausam & Etzioni O. (2010). Adapting open information extraction to domain-specific relations. AI Magazine, 31(3), pp. 93–102.
18. Wu, F & Weld, D.S. (2010). Open Information Extraction using Wikipedia. In proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL10), pp. 118-127.