International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:13, No:6, 2019

Discovering Semantic Links Between Synonyms, Hyponyms and Hypernyms

Ricardo Avila, Gabriel Lopes, Vania Vidal, Jose Macedo

***Abstract*—**This proposal aims for semantic enrichment between glossaries using the Simple Knowledge Organization System (SKOS) vocabulary to discover synonyms, hyponyms and hyperonyms semiautomatically, in Brazilian Portuguese, generating new semantic relationships based on WordNet. To evaluate the quality of this proposed model, experiments were performed by the use of two sets containing new relations, being one generated automatically and the other manually mapped by the domain expert. The applied evaluation metrics were precision, recall, f-score, and confidence interval. The results obtained demonstrate that the applied method in the field of Oil Production and Extraction (E&P) is effective, which suggests that it can be used to improve the quality of terminological mappings. The procedure, although adding complexity in its elaboration, can be reproduced in others domains.

***Keywords*—**Ontology matching, mapping enrichment, semantic web, linked data, SKOS.

I. Introduction

THE Semantic Web provides for the exchange and integration of data on the Web in a way that enables the iteration between humans and machines to be possible on a global scale. Ontologies are organic elements of the Semantic Web that are used to determine the knowledge exchanged or shared between different systems [9].

Interactive services such as Linked Data Mashups (LDM) are available on the Web by integrating the content of different data sources into new services [16]. The LDM thus makes it possible to create new services, providing mechanisms for the publication, the retrieval and the integration of data distributed throughout the Data Web [10], [5].

To generate new services and new knowledge, the Semantic Web and natural language worlds must be connected. These new knowledge structures create a bridge between the components of an ontology - classes, properties, and individuals - and their correspondents in natural language. However, the manual construction of ontologies is a difficult, tedious, erroneous and time-consuming process, usually requiring the collaboration of domain experts to validate the data model built to represent the set of concepts and the relationships between them.

We here present a new approach to the enrichment of ontologies that identifies different types of correspondences, such as synonyms, hyponyms and hyperonyms in the relationships between ontologies, within the oil domain, aiming at the integration of several glossaries of terms related to Oil Exploration and Production (E&P).

Ricardo Avila is with the Federal University of Ceara, Brazil (e-mail: ricardo.lims@gmail.com).

Other approaches have been proposed to identify such mappings (see Section IV), but they are still far from perfect. In general, previous studies have tried to directly identify different types of relationships, usually with the help of dictionaries such as WordNet. On the other hand, we propose an enrichment strategy using a two-step approach. In the first step, we apply the use of a crawler to collect synonyms, hyponyms, and hyperonyms to generate an XML file, then we generate new relationships between the instances and determine the initial ontology mapping with approximate matches of equality. Then, using a unified lexical ontology such as WordNet, we collect the lexical variants of each instance and their most likely type of relationship. Fig. 1 illustrates how the enrichment approach can improve mapping by identifying multiple relationships. As we will see in the evaluation, we can still achieve a high level of efficiency in the integration between heterogeneous bases.

The remainder of this article is organized as follows: Section II presents the proposed methodology; Section III shows the experiments and results; Section IV discusses related work, and finally; Section V outlines the conclusions and future work.

II. Proposed Approach

In order to generate new semantic links we use WordNet, which is a dictionary with lexical resources structured into groups of semantically related items that can be freely used as they are in the public domain. These resources are the main sources for the selection of semantically related lexicons in the field of interest of the present study, namely Oil Exploration and Production.

This proposal for the semi-automatic/automatic generation of synonyms, hyponyms and hyperonyms uses the skos:prefLabel predicate from the Simple Knowledge Organization System (SKOS), a data and vocabulary model that provides formal representation of knowledge representation structures, such as catalogues, glossaries, thesauri, taxonomies, folksonomies and other types of controlled vocabularies [15]. Three vocabularies were used in this study: The Glossary of the Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP), the Petroleum Dictionary in Portuguese (DPLP) and a set of oil concepts crawled from Wikipedia. The process for choosing these glossaries was described in the paper by [19].

A. Definition of Linguistic and Semantic Relations

An ontology O is a set of concepts C and its relation R, where each $r \in R$ connects two concepts $c_1, c_2 \in C$.

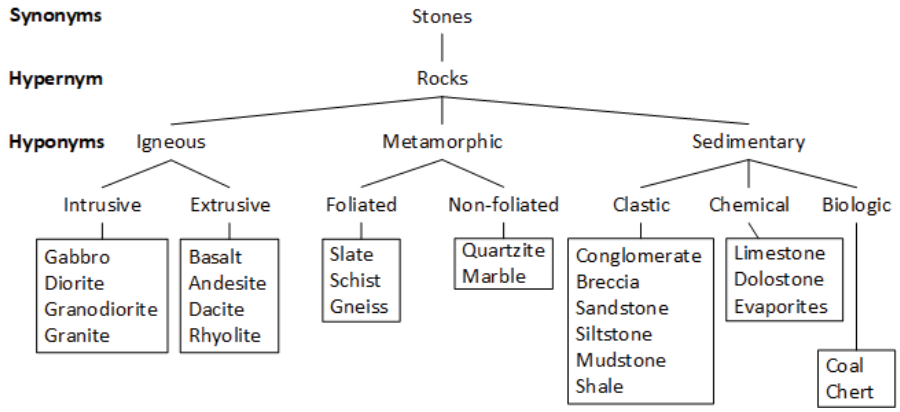

Fig. 1 Approach to discovery lexical types

A correspondence C between two ontologies O_1 and O_2 consists of a source concept $C_s \in O_1$, and a target concept $C_t \in O_2$. The relationship between two concepts c_1 and c_2 is formed by a similarity value between 0 and 1, expressing the computed probability of correspondence [17]. We have made use of the edit distance in strings, also known as the Levenshtein distance [13], to calculate the similarity between concepts, in conjunction with the lower-case textual pre-processing technique, since string comparison algorithms differentiate between upper and lower case.

Two concepts $c_1 \neq c_2$ of a language are called synonyms if they refer to the same semantic concept, that is, when they are similar or equivalent in meaning. For example, stone (pedra) and rock (rocha), according to WordNet in Portuguese, are words that mean the same thing.

Hyponyms show the relationship between a generic concept (hyperonym) and a specific instance of it (hyponym). A hyponym is a term or phrase whose semantic field is more specific than its hyperonym. c_1 is a hyperonym of c_2 if it describes something more general than c_2. c_2 is then called the hyponym of c_1. c_1 is a direct hyperonym of c_2 if there is no c_3 term, then that term c_3 is a hyperonym of c_2 and c_1 is a hyperonym of c_3.

Hyperonyms and hyponyms are asymmetrical. Hyponymy can be tested by substituting c_1 and c_2 in the sentence "c_1 is a type of c_2" and analyzing their veracity. For example, the sentence "A rock is a type of natural object" makes sense, but "A natural object is a type of rock" does not. Hyperonym is the inverse of hyponym.

B. Description of Vocabularies and External Resource

ANP is the Brazilian government's regulator of the oil, natural gas and biofuel industries. The ANP glossary has labels in Portuguese and contains 581 concepts that cover the main activities, processes, events, products, operations and other concepts related to the services of exploration and production of oil, natural gas and biofuels.

The DPLP is available on the Web and presents the standardization of technical concepts related to oil and gas, inherent to both research and production, as well as the regulatory and contractual aspects of this sector. It has 17,168 concepts, with labels in Portuguese, distributed across the areas of "Reservoir Technology", "Geology and Geophysics", "Production Technology", "Regulation and Contracts" and "Well Technology".

The set of concepts listed in Wikipedia totals 271 with Portuguese labels and all related to the production of oil and gas. In order to obtain the concepts from Wikipedia, a process of extraction, base persistence and, finally, of triplication was carried out.

The stage of collection and persistence of these vocabularies' concepts was done using a crawler. In the case of the triplication it was decided to develop a routine in the Python programming language for the execution of this process. This strategy was chosen due to the low maintenance effort of this type of routine, as well as the short learning curve and improvements in the processes of extraction, transformation and loading (ETL).

These three vocabularies make use of the SKOS vocabulary and data model in their representation. The glossaries are fully compatible with SKOS and consist of a set of skos:concept, where each of the concepts is identified by its respective skos predicate:prefLabel and skos:inScheme, in addition to having zero skos:altLabels, skos:definition, skos:hiddenLabel, skos:subject and skos:scopeNote. Furthermore, concepts can also be labeled with a zero or one dc:creator, dc:date, skos:topConceptOf, and skos:hasTopConceptOf and may have zero or more petro:urlImage, petro:legislation, petro:urlProvider and petro:referencedTerm. The three glossaries differ in size and coarseness, having, according to Fig. 2, the classes and relationships of the ontology used in this work.

WordNet is a unified lexical ontology for Portuguese. It was used as the main external source for the mapping of synonyms, hyponyms and hyperonyms. This source of knowledge has, in addition to the lexical variants used in this work, nouns, verbs, adjectives and adverbs. Currently, in its Portuguese version 3.0, it consists of 50,546 synsets (10,047 noun synsets), which were properly attributed by linguists.

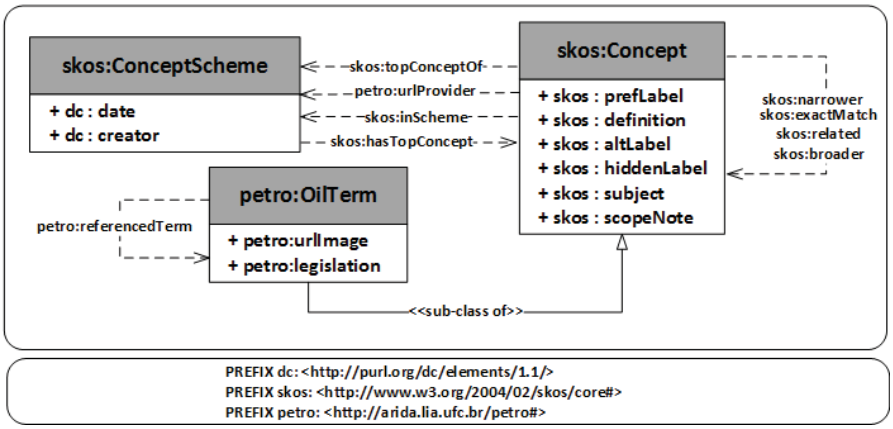

Fig. 2 Domain Ontology of the Glossary Mashup

C. Strategy Implemented

In the enrichment process of ontologies, there is no common terminology for the discovery of semantic relations. Some authors use language strategies, object orientation (such as UML), machine learning, or make use of other terminology types to identify the match types. Fig. 3 summarises the linguistic relationships, correspondence types, and predicates in RDF format that were generated.

Linguistic Relation	Example	Correspondence Type	RDF Type
Synonymy	Rock, Stone	equal, same_as	similarTo, has_synonym
Hyponymy	Rock, Natural Object	is_a, more_especific, less_general	hyponymOf, has_hyponymy
Hypernymy	Granite, Igneous Rock	inverse is_a, less_especific, more_general	hypernymOf, has_hypernymy

Fig. 3 Enrichment process to discover lexical types

This present work means to use the label of the predicate skos:prefLabel of the classes and properties of the ontology, identifying and retrieving the morphosyntactic characteristics of each concept and their respective synonyms, hyponyms and hyperonyms, converting them to RDF format and generating a lexical database with the relationships discovered. Fig. 4 shows the different steps of the process.

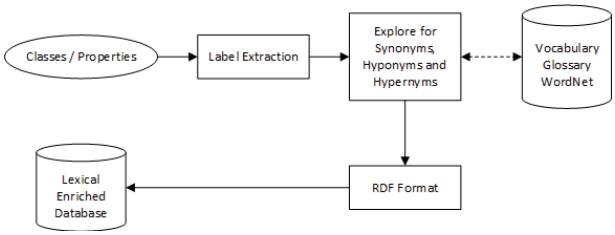

Fig. 4 Process construction flowchart

The approach adopted in this work makes use of one or more lexical entries, using glossaries of concepts related to Oil Exploration and Production (E&P) for each class and property. The first step involves the extraction of ontology labels and additional information. Then, if feasible, a search of synonyms, hyponyms and hyperonyms in external sources is carried out. The next step is to convert the semantic relationships into RDF (s, p, o). Finally, the generated triples are stored in a lexical database. Given an ontology O and an external resource R, both in the same language, new relations between the concepts $O(c_1) = R(c_2)$ are identified by following the steps:

- Retrieve all the synonyms of Oc_1 identified in $R(c_1)$, denoted as $S(c_1)$
- Generate a has_synonym relation of Oc_1 for each synonym of $S(c_1)$
- Retrieve all the hyponics of Oc_1 identified in $R(c_1)$, denoted as $H(c_1)$
- Generate a has_hyponym relation of Oc_1 for each hyponym of $H(c_1)$
- For each hyponym of $H(c_1)$, a has_hyponym relation with every synonym of $S(c_1)$ is generated
- Generate a has_hypernym relation for each hyponym of $H(c_1)$

Clearly this approach depends on the quality of the sources used to determine the match types for the concept pairs. The vocabularies and the external source (Section II.B) that were used are considered as being highly reliable. WordNet is a database used in several areas of Computational Linguistics and Language Engineering, such as automatic translation, search and information retrieval systems, information recovery systems, among others. It is organized according to the EuroWordNet general model, a multilingual database that supports the integration of WordNets from several European languages. The relationships generated after the execution of the five step terms are stored in a corpus for validation. This stage aims to identify any kinds of anomaly that might exist, such as concepts classified as homographs or perfect homonyms, which have the same writing, but with different meanings. In this type of situation, the description (definition) of the concept, its category, its domain or its application and use can be verified, depending on the type of information that the source used provides. As the WordNets cover several domains of knowledge, this kind of validation is very important to avoid absurd errors, such as, for example creating relationships between the Portuguese concepts of corte (a

court) and corte (a cut), or similarly the concept ”bow”, in English, which also has many very different applications. As in natural languages it is common to use more than one word to convey the same meaning, one of the objectives of this work was to identify as many synonyms as possible for the concepts of the ontologies. In order to address the idea of polysemic concepts and to collect those that are most relevant within the applied domain, the [12] approach was used here.

III. Experiments and Results

This section explains the metrics that were used to validate and evaluate the accuracy and efficiency of the proposed ontology enrichment method. Ontologies were used in the field of Oil Exploration and Production (E&P) in our experiments. The proposed method uses a lexical database (WordNet) at the center of the process to discover new relationships between concepts.

If one uses manual mapping to discover new relationships, perfect or nearly perfect results can be acheived. Therefore, we want to determine the general quality of the relationships as well as the quality of the lexical variants discovered. It is worth noting that the level of representational coarseness of the vocabularies and external resource utilized can generate distinctions within the context of a specific domain. To mitigate this type of issue, we introduced different assessment measures.

In the remainder of this section, we first present the measures used to assess the quality of relationships (Section III.A). Then, we evaluate the method of discovering new relationships (Section III.B) using two sets of suchnew relationships, one being generated automatically, and the other being mapped manually by a domain expert. Finally, we apply the metrics of precision, coverage, F-measure and confidence interval to evaluate the quality of the lexical variants discovered (Section III.C).

A. Evaluation Metrics

Precision, coverage and F-measure were employed to evaluate the number of concepts and their respective discovered synonyms, hyponyms and hyperonyms, for each ontology used in the experiments. In this way, we were able to prove the quality of the vocabularies independent of the level of representational coarseness of each one.

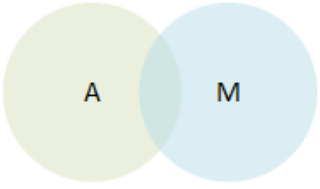

Fig. 5 Automatic and Manual Mappings

Fig. 5 shows two sets of relations, A being generated automatically, and M being mapped manually by a domain expert. We determined the precision (p), coverage (r) and F-measure (f), based on the number of relationships identified for each concept:

$$p = \frac{|A \cap M|}{M},\ r = \frac{|M \cap A|}{A},\ f = \frac{2pr}{p+r}$$

To validate the quality of the lexical variants discovered, 0.10% of the total new discoveries were selected. The new relationships were selected following an arbitrated formula with the establishment of a base and an index. A base 2 and an index 2 were used which resulted in the selection of the even register lines in the files. Sampling populations were calculated based on a 95% confidence level and a confidence interval of 5 [6].

B. Evaluating Automatic and Manual Mappings

At this evaluation stage, the vocabularies and external sources detailed in Section II.B were used. For each vocabulary a process of finding synonyms, hyponyms and hyperonyms was applied, generating two sets of relations, with A generated automatically and M mapped manually by a domain expert. Fig. 6 provides relevant information on the discovery tasks of lexical variants.

Vocabulary	Concepts	Synonyms		Hyponyms		Hypernyms	
		A	M	A	M	A	M
ANP	581	278	132	332	119	610	251
DPLP	17.168	6.872	5.224	10.235	4.754	17.107	9.978
Wikipedia	271	96	122	188	115	284	237

Fig. 6 Number of lexical variants discovered

In the first mapping scenario (ANP v WordNet), following the automatic strategy, 47.85% of synonyms and 57.14% of hyponyms were discovered in relation to the total number of concepts (581). According to the proposed approach, 610 hyperonyms were created, which is exactly the sum of the synonyms and hyponyms discovered. This amount is due to step five of the strategy that was implemented (Section II.C) and the asymmetric characteristic between hyponyms and hyperonyms. Manually mapped relationships were 22.72% and 20.45%, respectively for synonyms and hyponyms, with 251 hyperonyms found in all.

In the following scenario (DPLP v WordNet), 40.03% of synonyms and 59.62% of hyponyms were discovered, out of a total of 17,168 concepts, in addition to the creation of 17,107 hyperonyms. This was the best mapping scenario and evidence of vocabulary quality and the automatic relationship discovery strategy, since the manual approach obtained 30.43% of synonyms and 27.69% of hyponyms, with the generation of 9,978 hyperonyms in all.

Finally, in the last scenario (Wikipedia v WordNet), the automatic discovery of 34.42% of synonyms, 69.37% of hyponyms and the generation of 284 hyperonyms occurred. On the other hand, the manual approach obtained better values in this scenario, with the discovery of 45.02% of synonyms, 42,32% of hyponyms and the creation of 237 hyperonyms in total.

The results of the automatic mapping scenarios were considered relevant, since they involved the discovery of semantic connections using a more general domain lexical

ontology. The three ontologies largely covered the same domain, namely that of Oil Exploration and Production (E&P).

C. Evaluation of the Quality of Lexical Variants

This is considered the main assessment of the proposed approach. The automatic identification of lexical variants obtained a large number of correspondences in the scenarios proposed in Section III.B, but only the true positives can be considered relevant. Table 7 presents the results of precision (p), coverage (r) and the F-measure (f) after the validation of the semantic links generated by a domain expert. It is worth noting that only the results of the automatic discoveries are presented.

	r	p	f
ANP	.64	.99	.77
DPLP	.87	.92	.90
Wikipedia	.48	.70	.57

(a) synonyms

	r	p	f
ANP	.90	.57	.70
DPLP	.96	.89	.92
Wikipedia	.66	.90	.77

(b) hyponyms

	r	p	f
ANP	.88	.88	.88
DPLP	.93	.93	.93
Wikipedia	.76	.76	.76

(c) hypernyms

Fig. 7 Evaluation of scenarios

The discovery of synonyms obtained an F-measure of between 57 and 90%, indicating a good level of efficacy in the proposed approach. The accuracy was also good (70 to 99%), although the coverage was slightly limited in the case of the Wikipedia (48%). Given the importance of generating synonyms, we consider that the results were relevant as the approach used ultimately depends on the quality of the sources it uses to determine the match types for the concept pairs.

The hyponyc discovery scenario, however produced better results, with the F-measure varying in its values between 70 and 92%, the precision between 57 and 90% and the coverage between 66 and 96%. These values were most likely influenced by the more generalist comprehension of WordNet and, mainly, by the step of identifying anomalies.

Finally, hyperonyms, which are generated based on synonyms and hyponyms, obtained the same values for coverage, precision and the F-measure, with results of between 76 and 93%, proving the quality of the relationships discovered. This proves that the automatic discovery of the lexical variants proposed in this work works very well if the false positives are disregarded.

IV. Related Work

There are a number of studies that try to identify different types of relationships for semantic enrichment [2], [1], [8], [18]. Lexical repositories such as WordNet [3] and Cye [7] are high-quality sources, but with low coverage of concepts and instances. Manual construction of ontologies is a difficult process, usually requiring expert collaboration. Automatic methods for constructing and enriching ontologies have emerged to reduce the effort and facilitate this process. Our approach uses a unified lexical ontology, such as WordNet in Portuguese, and glossaries in the oil domian to collect lexical variants of each instance and their most likely type of relationship, creating new instances of synonyms, hyponymous and hyperonimous in vocabularies from the domain of Oil Exploration and Production (E&P).

The work of [4] presented a new structure similarity algorithm, using fuzzy logic, for the automatic alignment of abbreviated words and synonyms. The diffuse logic algorithm employs a new diffuse punctuation method to increase the accuracy of the proposed procedure. The results showed that precision quality was increased by 10% in each test set. The approach in this work presents tests with short lists and synonyms, being one of the techniques used in the validation process during the research stages of this article.

The work by [11] proposes an alignment approach using association rules. They use the YAGO, DBpedia and Freebase vocabularies to discover sameAs links, among other relations. Even though this work does not deal with the identification of lexical variants, its experiments use knowledge bases in different domains and a strategy of automatic discovery of corresponding relationships in the target data set, which is similar to the scenarios of this present work.

An automatic method for the enrichment of ontologies in the biomedical domain is proposed by [14], where the lexical/syntactic complexity, the semantics and the extraction of biomedical terms from a specialized text corpus are discussed. The results obtained from the discovery of new predicates, taking into account the paradigmatic relations, that is, synonyms, hyperonyms (parents) and hyponyms (children). This work is very similar to our study but is applied to a different domain and uses natural language processing for the automatic discovery of new relationships.

V. Conclusions and Future Work

This article offers a new approach to the semantic enrichment of ontologies for the automatic discovery of synonyms, hyponyms and hyperonyms in the domain of Oil Exploration and Production (E&P). We use the skos:prefLabel predicate label of ontology classes and properties to identify and retrieve the morphosyntactic characteristics of each concept and its lexical variants, converting the discovered relationships into RDF format and storing them in a lexical database. To acheive an alignment of the concepts, the Levenshtein algorithm was used to calculate similarity in conjunction with the lowerCase textual preprocessing technique. Even using more specific vocabularies and WordNet, a more general repository covering different areas of knowledge, our approach presented satisfactory results. This technique can be applied to other domains and languages aside from Portuguese, requiring only some minor adjustments.

As a suggestion for future work in this area, we would recommend the use of vocabularies in other domains and areas of knowledge, as well as the use of other thesauri and reliable sources. In addition, other methods of alignment between classes/properties could be used as rules of inference or based on supervised automatic learning. The use of machine learning techniques are also recommended to identify new semantic relations, predicates, classes and properties through the use of the description of instances (skos:definition).

Our research will continue because we understand that there is a need to improve the results of lexical variant discoveries in other languages and to further develop strategies to identify anomalies in the cases of perfect homonyms or homonyms.

REFERENCES

[1] David Aumueller, Hong-Hai Do, Sabine Massmann, and Erhard Rahm. 2005. *Schema and ontology matching with COMA++*. InProc. ACM SIGMOD.

[2] Jiaqing Liang, Yi Zhang, Yanghua Xiao, Haixun Wang, Wei Wang, and PinpinZhu. 2017. *On the Transitivity of Hypernym-Hyponym Relations in Data-Driven Lexical Taxonomies.* In AAAI. AAAI Press, 1185–1191.

[3] George A. Miller. 1995. *WordNet: a Lexical Database for English.Commun.* ACM38, 11 (1995), 39–41.

[4] Akarajit Tanjana and Jian Qu. [n. d.]. *Enabling Fuzzy Logic to Enhance Auto-matic Schema Matching.* In Science & Technology Asia, Vol. 22. 93–111.

[5] Christian Bizer, Richard Cyganiak, and Tobias Gauss. 2007. *The RDF BookMashup: From Web APIs to a Web of Data.* In3rd Workshop on Scripting forthe Semantic Web.

[6] J. M. Bland and D. G. Altman. 1996. *Transformations, means, and confidenceintervals.* BMJ312, 7038 (April 1996), 1079.

[7] D. Lenat and R. V. Guha. 1990. *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project.* Addison-Wesley.

[8] William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. 2003. *A Com-parison of String Distance Metrics for Name-Matching Tasks.* In Proceedings of the IJCAI-2003 Workshop on. Acapulco, Mexico, 73–78.

[9] Mathieu d'Aquin, Gabriel Kronberger, and Mari Carmen Suárez-Figueroa. 2012. *Combining Data Mining and Ontology Engineering to Enrich Ontologies and Linked Data.* InKNOW@LOD, Vol. 868. CEUR-WS.org, 19–24.

[10] Tom Heath and Christian Bizer. 2011. *Linked Data: Evolving the Web into a Global Data Space (1st ed.)*. Morgan Claypool.

[11] Maria Koutraki, Nicoleta Preda, and Dan Vodislav. 2016. *SOFYA: Semanticon-the-fly Relation Alignment.* In EDBT. OpenProceedings.org, 690–691.

[12] M. Lesk. 1986. *Automatic sense disambiguation using machine readabledictionaries: how to tell a pine cone from an ice cream cone.* In Proceedings of ACM SIGDOC Conference. Toronto, Ontario, 24–26.

[13] V. Levenshtein. 1966. *Binary Codes Capable of Correcting Deletions, Insertions, and Reversals.* Soviet Physics-Doklady10, 8 (1966), 707–710.

[14] Juan Antonio Lossio-Ventura, Mathieu Roche, Clément Jonquet, and Maguelonne Teisseire. 2016. *A Way to Automatically Enrich Biomedical Ontologies.* In EDBT, OpenProceedings.org, 676–677.

[15] Alistair Miles, Brian Matthews, Michael Wilson, and Dan Brickley. 2005. *SKOScore: Simple Knowledge Organisation for the Web.* In Proc. of the 2005 international conference on DC and metadata applications. Madrid, Spanien, 1–9.

[16] Markus Nentwig, Michael Hartung, Axel-Cyrille Ngonga Ngomo, and Erhard Rahm. 2015. *A survey of current Link Discovery frameworks.* Semantic WebPreprint (2015), 1–18.

[17] Jacco van Ossenbruggen, Michiel Hildebrand, and Viktor de Boer. 2011. *Interactive Vocabulary Alignment.* InTPDL (Lecture Notes in Computer Science), Vol. 6966. Springer, 296–307.

[18] Ziqi Zhang, Anna Lisa Gentile, Eva Blomqvist, Isabelle Augenstein, and FabioCiravegna. 2013. *Statistical Knowledge Patterns: Identifying Synonymous Re-lations in Large Linked Datasets.* In International Semantic Web Conference(1), Vol. 8218. Springer, 703–719.

[19] Ricardo Ávila, Salomão Santos, David Araújo, Vânia Maria Ponte Vidal, and José Antônio Fernandes de Macêdo. 2017. *Semantic Links Using SKOS Predicates.* In KES (Procedia Computer Science), Vol. 112. Elsevier, 467–473.