ISSN 1870-4069

Semantic Integration of COPEs in GOLD Ontology

Malek Lhioui[1], Kais Haddar[1], Laurent Romary[2]

[1] Sfax University,
Laboratory MIRACL, Multimedia, Information Systems and Advanced Computing Laboratory,
Tunisia

[2] Inria & Centre Marc Bloch,
Germany

laurent.romary@inria.fr, ma.lhioui@gmail.com,
kais.haddar@yahoo.fr

Abstract. This paper has as goal the semantic integration of the local ontologies named COPEs (Community of Practice Extension) in GOLD (General Ontology for Linguistic Description) ontology. COPEs are OWL ontologies that include specific knowledge. However, GOLD is a global ontology that includes general knowledge of the field. Thus, we deal with the challenge of the construction of a global schema presented as global ontology for lexical resources introduced as local ontologies. The originality of our suggestion consists on the use of a web semantic method to resolve an NLP issue, i.e. semantic integration as method for interoperability resolution between lexical resources. We define lexicons as local ontologies using Description Logics. Then, we build the resulted global ontology by combining alignment techniques and logical-reasoning.

Keywords: Semantic integration, COPEs, GOLD, interoperability.

1 Introduction

As the need for interoperability between lexical resources is increasing, we deal with the challenge of building a global schema for lexical resources. Indeed, several linguistic consortiums, such as ISO, need to share particular fragments of specific domains. Users will not be obliged to learn about the details of many sources like structures, vocabularies, concepts, relations, etc. Consequently, users are conciliated when formulating queries. There are several formalisms of lexical resources such LMF, TEI, HPSG, etc. However, formalisms may dispose of incomplete and partial data because some of them do not offer full knowledge. For example, TEI does not offer complete semantic information.

For this purpose, linguists need to exchange information between lexical resources. The exchange of information seems to be necessary to provide a more complete linguistic resource which reply to users' needs.

adfa, p. 1, 2011.
pp. 71–81; rec. 2017-02-08; acc. 2017-04-12

The method presented here allows constructing a global schema: ontology treated as pivot format for lexical resources. For this construction, we use semantic integration because of its large interest dealing with heterogeneous knowledge, ensuring enrichment and avoiding destruction of old editions.

Building such a global ontology for reasons of interoperability between lexical resources may face several problems. First, we have to perfectly choose the optimal representation language for our case (RDF(S), OWL-Lite and OWL-DL). Besides, the construction of such ontology requires a precise knowledge of the lexical resources, their heterogeneity in the distribution of knowledge, the used nomenclature and their coverages (lexical, syntactic, etc.). From a technical point of view, the collection of lexical resources is a big dilemma for the majority of specialists in the language community. In our case, we use ALIF platform so we need to learn carefully about it.

Our method is original since there are no previous operable works trying to resolve interoperability between lexical resources. Moreover, using ontologies for resolving a big issue which is interoperability between lexical resources is in itself an innovation. From another point of view, interoperability nowadays become a big issue and recent projects must take care of it otherwise they are out of progress. The [1] report states that: “The lack of interoperability costs the translation industry a fortune”. Fortune is compensated to regulate the adjustment of lexical resources.

The method we already introduce in this paper is operable whatever the language. In the following parts of the paper, we give a concise state of the art talking about our big topic which is interoperability between lexical resources. Then, we define prerequisites required in introducing our method. Then, we define our proper new method which is using semantic integration between local ontologies to build global one. Finally, we bring to a close with experimentation and evaluation section.

2 Related Works

There are no previous works dealing with the use of semantic integration for the purpose of resolving interoperability between lexical resources. However, there are works related with semantic integration and others concerning interoperability between lexical resources. Since there are two main separated topics, we classify the following state of the art into two main parts. The first part concerns the semantic integration. In the second part, we discuss interoperability between existing lexical resources.

2.1 Semantic Integration

Semantic integration is a recent approach which is based on ontology integration. In order to formalize this approach, experts use Description Logics and ontology alignment. In [2], authors present a distributed description logic.

The formalism presented in [3] defines aspects of distributed and modular ontology reasoning. In the two cases, authors try to define that the concept of distributed description logics is needed for relating various data sources. In [4], authors introduce a new approach “E-connections framework” as a solution for connecting different sources. The defined approach is.

2.2 Interoperability between Lexical Resources

Since there are no serious cited efforts in literature aiming to resolve interoperability between lexical resources in NLP domain, we discuss the bidirectional mapping between formats of lexical resources. [6] is the first mapping process converting HPSG lexicons to OWL ontology. A rule-based system is invented by [7] in order to translate LMF syntactic lexicon into TDL using the LKB platform. Then, a prototype for projection HPSG syntactic lexica towards LMF have been developed by [8]. In the same context, a mapping process converting LMF lexicons to OWL ontologies is described in [9].

These works usually involve two formalisms, processing more than two formats is a hard task even impossible. In order to appease the difficulty of transformation process, the ISO try to solve the problem with a normalization process. It proposes an ISO standard in 2003named LMF [10]. All these works are deeply linked to our proposed method. However, we use the approach of semantic integration to introduce a new method for resolving interoperability between lexical resources. In the following part, we familiarize with notions required to define our method.

3 Prerequisites

We use for semantic integration reference ontology: on the one hand links between source ontologies are obtained from the taxonomical relationships of the reference ontology. On the other hand, mappings between the global ontology and sources are obtained by syntactic-matching, from source-concepts names to reference-ontology-concepts names.

3.1 Ontology of Reference: GOLD

GOLD[1] (General Ontology for Linguistic Description) is a general ontology described in OWL including linguistic knowledge as well as a qualified linguist [12]. Knowledge including in this ontology consists on the core of any theoretical framework. Furthermore, GOLD incorporates knowledge concerning descriptive linguistics. For example, “an adjective is a part of speech” [12]. Linguistics communities consider GOLD as a reference ontology that uses language-neutral and theory-neutral terminology. For example, LexicalResource is a subclass of gold: Entity.

3.2 Local Ontologies: COPEs

Local ontologies are sub-communities of practice considered as instantiations noted COPEs (Communities of Practice Extension (COPEs). COPEs are simple OWL ontologies that import local knowledge to a global resource [11]. In order to give a real example of components in COPE of LMF, we note the part of speech propriety designed as subclasses of lmf: pos. Parts of speech in LMF have perhaps their

[1] http://www.linguistics-ontology.org

equivalent in GOLD ontology. The liaison will be done automatically by means of semantic integration. This phenomenon is described clearly in the next section.

3.3 Semantic Integration

Data structures whatever their kinds (non-structured, semi-structured and structured) are more and more complex. Therefore, their handling is no longer simple. Indeed, data present an accessibility issue because of its different kinds. However, though their heterogeneity, several data sources are semantically related. Different concepts describe the same reality. The access to these data constitutes a big dilemma because of the non-precision of their localization. Consequently, interoperability is so required in this case between a new created system playing the role of interface and the other sources. The new created system is based on a data integration process offering a new interface for distributed, heterogeneous and independent sources.

4 Semantic Integration of COPEs in GOLD Ontology

Data semantic integration is in general progress with the evolution of data structures (XML, RDF, OWL, etc.). The goal of our proposed paper is to build global ontology from local ontologies using of reference ontology GOLD using semantic integration. In order to formalize our domain, we exploit description logics DL to define a domain by a set of:

– Concepts: which express classes and manipulate them as objects, example: Lexical Entry, Lemma, Stem, etc.
– Roles: which express relations and operate them as relations between objects, example: Lexical Entry related To Lemma.

An ontology $O = < T, A >$ is composed of:

– $TBox\ T(intensional\ Knowledge)$: Identify general propriety of concepts and roles,

In order to illustrate ontology components, we give the following example:

$\exists\ relatedTo\ LexicalEntry$

$\exists\ relatedTo\ Lemma$

$\exists\ hasForm\ LemmatisedForm$

$\exists\ hasForm$ InflectedForm

$InflectedForm \subseteq LemmatisedForm$

$InflectedForm \subseteq \delta(writtenForm)$

$InflectedForm \subseteq \delta(number)$

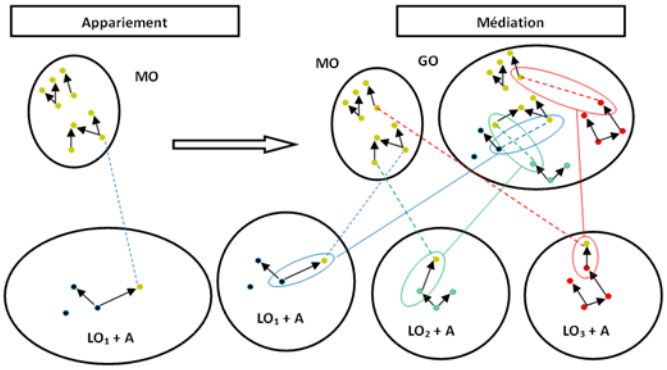
$\rho(writtenForm) \subseteq \delta(xsd{:}\,string)$

$\rho(number) \subseteq \delta(xsd{:}\,string)$

– $ABox\ A\ (extensional\ knowledge)$: Identify assertions related to concepts and roles instances.

In order to more explain the composition of the ontology, we give a real example:

– $InflectedForm(inflectedForm)$
$writtenForm(inflectedForm, clergymen)$
$number(inflectedForm, plural)$



MO: Mediator ontology
GO: Global ontology
LO_i: Local ontology
A: Anchoring

Fig. 1. Proposed method.

After presenting DL-ontologies, we define the required foundations in our proposal. Here are three mains needed $TBox$es in our proposed method:

– A set of local $TBox$ called$\{Tli\}$: They present involved local data sources {Si} in the sharing process.

– A $TBox\ Tm$: It provides intentional knowledge extracted from the ontology of reference.

In order to more explain the composition of the ontology, we give a real example:

– A Tg : conciliates the different local $TBox$ and supplies a shared conceptual level of the domain application.

After this formalization, the issue of construction a global ontology is summarized in the build of $TBox\ Tg$. This Tg integrates $TBox$ of local ontologies and adjusts their concepts using $TBox$ of the reference mediator.

The previous fig.1 shows the two main steps of the proposed method: appariement and mediation. The method is based on the use of automatic reasoning functions of description logics in order to automate the construction process. Then, we use ontology of reference to have an appropriate conceptualization of the application domain. The ontology of reference has been developed independently from any specific objective by experts in knowledge and domain engineering: GOLD.

4.1 Appariement

The appariement step is based on modifications made in $Tl\ TBoxes$. Then, we integrate them in global $Tg\ TBox$. Fig. 2 shows the appariement step. Two sub steps have been to achieve: *anchoring process* and *automatic updating of local* $Tl\ TBox$ into Tla. This step allows linking concepts of Tl (anchored concepts) and concepts of referene mediator $TBox\ Tm$ (anchor concepts).

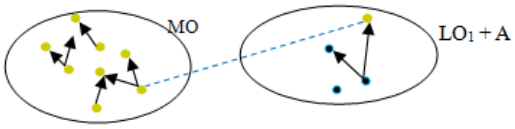


Fig. 2. Appariement of concepts between the mediator and the local ontology.

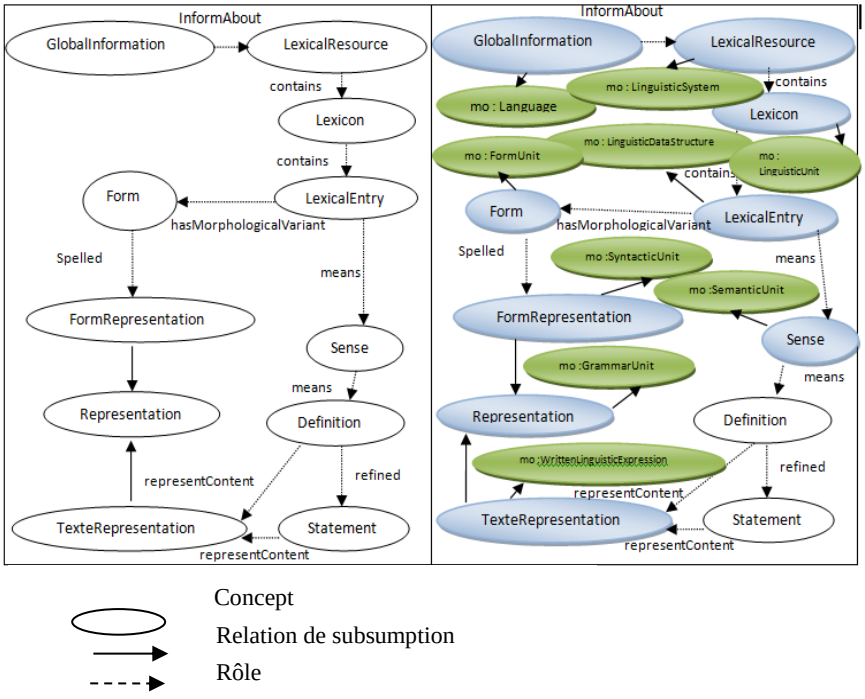


Fig. 3. Concrete example of the appariement step: application in LMF core model.

Fig. 2 summarizes the whole process of appariement step: anchoring and updating. In the anchoring process, we make liaison between concepts of the mediator (MO) and those of local ontology (LO). Fig. 3 shows a real example of the appariement step.

In fig. 3, two types of anchoring are established: lexical and semantic anchoring. For the first anchoring: It is simply matchings of Tl to Tm :

– Calculation of a set of mappings noted $M = \{mi\} \ / \ mi = Al \sqsubseteq Am \ \ (Al \in Tl, Am \in Tm)$.

– Using of lexical similarity measure δ: $\lceil NL \times \lceil Nm \rightarrow [0, 1]$ ($\lceil NL, \lceil Nm$ are a set of atomic concepts names.

Semantic anchoring: It is a question of finding additional anchoring concepts which are subsumed by anchored process.

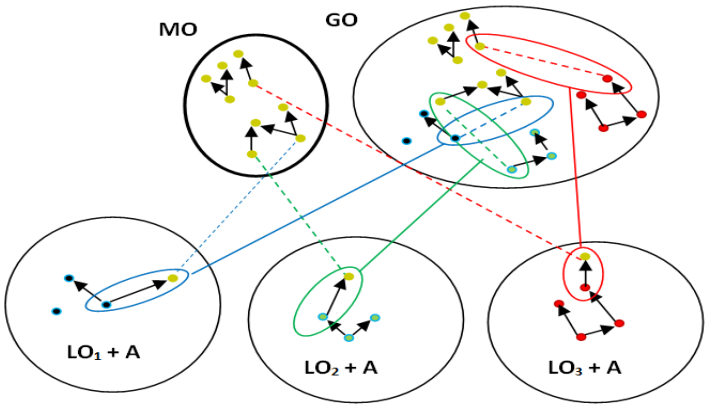


Fig. 4. Global procedure of the mediation.

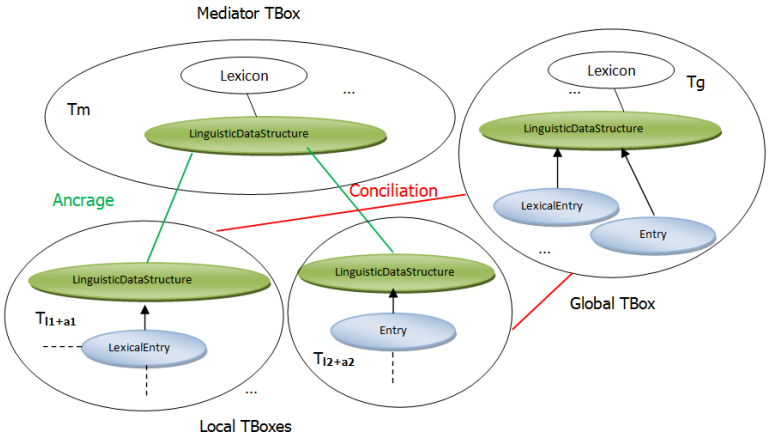


Fig. 5. Real example of mediation process.

After the anchoring process, we build a new Tla containing the result of the appariement:

We note $Tla =< Tl, M >$ the appariement of Tl compared to Tm.

– Tl is a local ontology.

– M is the result of anchoring Tl compared to Tm.

4.2 Mediation

The mediation consists on the integration process based on *related concepts computation for* Tla. The mediation step allows the construction of the global TBox Tg using the result of the appariement phase. The new ontology is reached *incrementally*

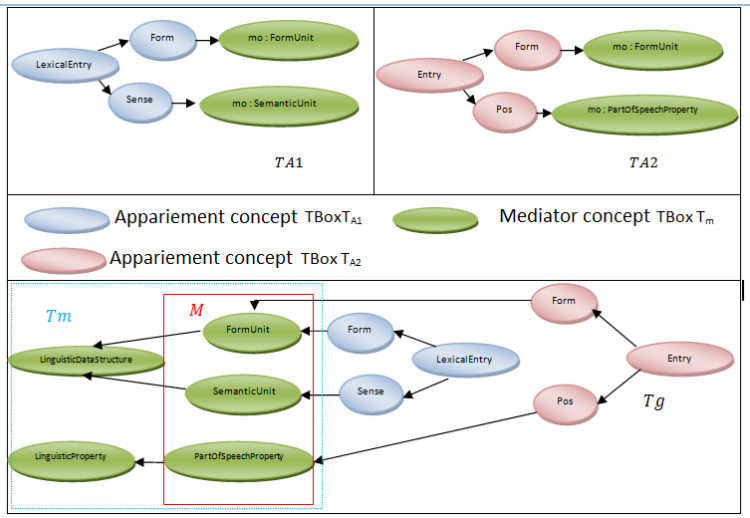


Fig. 6. Illustration of mediation algorithm.

by integrating the resulted TBoxes of the appariement step. It is a question of linking appariement TBoxes concepts. Fig. 4 illustrates the whole process of the mediation.

In fig. 4, the global ontology GO is incrementally built in two main sub-steps: *integrate local ontology and their appariement Tla* and *calculate all related subsumers for the anchor concepts*.

The first sub-step is summarized in the liaison between the resulted TBoxes from the appariement step and those of the mediator ontology. Concepts in the global ontology GO are linked incrementally with anchor concepts in the mediator ontology MO. Thus, the TBox Tg includes the following sets: - the set of appariement TBoxes *Tla*, and – a subset of Tm containing the related part of the hierarchy related with anchored concepts.

In order to give a real example between LMF and TEI ontology, we take the example in fig. 5. In this example the concepts *LexicalEntry* of the appariement TBox Tl_1+a_1 and *Entry* of the appariement TBox Tl_2+a_2 are respectively anchored by the same concept of the mediator TBox Tm: *LinguisticDataStructure*.

The structure of the mediator TBox reveals that *LinguisticDataStructure* has related concept. We add this relation to merge the concepts LexicalEntry and Entry in the global TBox Tg.

The fig. 5 can be generalized by the following algorithm of mediation:

- Input: $\{T_{Ai} = < T_{li+ai}, Mi >\}$, Tm
- Output: $Tg = < \{T_{Ai}\}, Tm >$

Each anchor concept Am of Mi is integrated in the Tg hierarchy.

Calculation of related concepts of Am in Tm called Arc (All Related Concepts) among present concepts in the hierarchy Tm

$Arc \leftarrow All\_related_{Tm}\ (Am)$

The previous algorithm is illustrated as follow:

$$T_{A1} = < T_{l1+a1}, M1 >$$

$$M1 = \{\text{Form} \subseteq \text{mo : FormUnit, Sense} \subseteq \text{mo : SemanticUnit}\}$$

$$T_{A2} = < T_{l2+a2}, M2 >$$

$$M2 = \{\text{Entry} \subseteq \text{mo : LinguisticDataStructure, Form} \subseteq \text{mo : Form}\}$$

In fig.6, we illustrate the previous algorithm of mediation using reference ontology GOLD as mediator (concepts colored with green) and logical-inference of description logics for automatic construction of the global ontology.

5 Experimentations and Evaluations

In order to evaluate our approach, a system of request is established using a construction process of requests exploiting special types of mappings. The build system is able to ask the appropriate local ontologies. The system interrogates the concerned sources and recomposes partial responses and restitutes the global response. This system is set up in a separated module but not explained in this paper.

After the description of the approach aiming to build automatically a global ontology, we introduce the different experimentations done in this section. In order to verify that our approach is feasible, we have to experiment it in order to prove that the approach allows producing a global ontology providing a shared conceptualization for involved sources with reasonable costs in terms of time and human resources. Our approach provides a facility of updating sources.

In order to evaluate our proposal, we choose a full ontology of reference in the lexical resources' domain. We realize the experimentations with the following techniques characteristics: a machine endows with a system Windows 7 with an Intel(R) Core (TM) processor. For the appariement step, we have used a well-known measure of lexical similarity called "string metrics" proposed by [13]. All local used ontologies are built automatically using an MDA[2] process [14]. The following notations used in the experimentation part are:

- LM: Lexical Mappings found after a lexical appariement, additional notations are given such 1;1 and 1;m used respectively to design mappings for one appariement found for one concept and mappings found for several candidates found for the same concept,
- VM: Validate Mappings,
- SM: Mappings found after a semantic appariement.

The table 1.1 illustrates results obtained after the appariement step for the two ontologies: LMF core ontology and TEI ontology. The two cited ontology are built automatically in [14] via the MDA approach.

[2] MDA : Model Driven Architecture

Table 1. Results of the appariement process in the context of LMF and TEI ontologies.

Ontologies	Concepts	Appariement							
		LM		VM	SM	Concepts selected for appariement			Time for execution
		1;1	1;m			pairings	selected	matching	
TEI Ontology	207	203	1	197	9	206	1	178	02 :39s
LMF Ontology	39	37	1	34	5	39	0	17	00 :49s

Table 2. Results of the mediation process in the context of LMF and TEI ontologies.

Pairings	Concepts for matchings	Mediation		
		Global ontology		Mediation time
		Local concepts	Concepts given by the reference ontology	
Ta1	178	207	83	05 :03s
Ta2	17	175	104	02 :01s

When analyzing results shown by the given table, we note that almost lexical mappings are 1;1. Among these found mappings, only 197 ones are valid. We note that many concepts selected for appariement share the same pairing: we have 206 pairings with 178 matchings. We remark that the time for execution reflects complexity of the construction process.

According to the results given by the table 2, the mediation time is not proportional to the concepts given because it depends on the process integration and not the number of concepts found for mediation. In fact, the integration is the task that consumes the time. Thus, the complexity is linear to mediation process.

6 Conclusion

In this proposal, we have proposed a new method for interoperability between lexical resources using semantic integration of lexical local ontologies in GOLD. Our method does not depend on the quality of involved lexical resources. The established method combines the results of MDA approach and ontology alignment techniques to make semantic links between involved lexical resources. Our research field contains four main parts: lexical resources, interoperability issue, semantic integration and ontologies alignment.

First of all, we have made a smart research on the main existing lexical resources in several languages. Then, we have made a great study on interoperability issue, and since there are no serious attempts to resolve this notion in NLP area, we have discussed the bidirectional mapping from one format to another. In future works, we have to extend our method using other metrics of the two-alignment ontology. In fact, interoperability appears to be solved since it combines two main approaches MDA Transformation and ontology alignment.

References

1. TAUS: Report on a TAUS research about translation interoperability (2011)
2. Borgida, A., Serafini, L.: Distributed description logics: Assimilating information from peer sources. Journal on Data Semantics I, Springer, vol. 2800, pp. 153–184 (2003)
3. Serafini, L., Borgida, A., Tamilin, A.: Aspects of distributed and modular ontology reasoning. IJCAI, pp. 570–575 (2005)
4. Grau, B. C., Parsia, B., Sirin, E.: Combining OWL ontologies using E-connections. Journal Web Semantics, vol. 4, no. 1, pp. 40–59 (2006) doi: 10.1013/j.websem.2005.09.010
5. Niang, C., Bouchou, B., Lo, M., Sam, Y.: Automatic building of an appropriate global ontology. In: East European Conference on Advances in Databases and Information Systems, pp. 429–443 (2011)
6. Wilcock, G.: An OWL ontology for HPSG. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pp 169–172 (2007)
7. Loukil, N., Ktari, R., Haddar, K., Benhamadou, A.: A normalized syntactic lexicon for arabic verbs and its evaluation within the LKB platform. ACSE (2010)
8. Haddar, K., Fehri, H., Romary, L.: A prototype for projecting HPSG syntactic lexica towards LMF, JLCL (2012)
9. Lhioui, M., Haddar, K., Romary, L: A prototype for projecting LMF lexica towards OWL, CICLING (2015)
10. Francopoulo, G.: LMF lexical markup framework. John Wiley & Sons, Inc (2013)
11. Farrar, S., Lewis, W. D.: The gold community of practice: An infrastructure for linguistic data on the web (2005) http://www.u.arizona.edu/˜farrar/
12. Stoilos, G., Stamou, G., Kollias, S.: A string metric for ontology alignment. In: Proceedings of the 4rd International Semantic Web Conference (ISWC), Lecture Notes in Computer Science, Springer, vol. 3729, pp. 624–637 (2005)
13. Lhioui, M., Haddar, K., Romary, L.: A new method for interoperability between lexical resources using MDA approach. Advanced Intelligent Systems and Informatics (2016)