Challenges for the Multilingual Web of Data

Jorge Gracia [a], Elena Montiel-Ponsoda [a], Philipp Cimiano [b], Asunción Gómez-Pérez [a], Paul Buitelaar [c], John McCrae [b]

[a] Ontology Engineering Group, Universidad Politécnica de Madrid
Campus de Montegacedo s/n, Boadilla del Monte 28660, Madrid, Spain
{jgracia, emontiel, asun}@fi.upm.es

[b] Semantic Computing Group, CITEC, University of Bielefeld
Universitätsstraße 25, D-33615 Bielefeld, Germany
{cimiano, jmccrae}@cit-ec.uni-bielefeld.de

[c] Unit for Natural Language Processing, DERI, National University of Ireland, Galway
IDA Business Park, Lower Dangan, Galway, Ireland
paul.buitelaar@deri.org

Abstract

The Web has witnessed an enormous growth in the amount of semantic information published in recent years. This growth has been stimulated to a large extent by the emergence of Linked Data. Although this brings us a big step closer to the vision of a Semantic Web, it also raises new issues such as the need for dealing with information expressed in different natural languages. Indeed, although the Web of Data can contain any kind of information in any language, it still lacks explicit mechanisms to automatically reconcile such information when it is expressed in different languages. This leads to situations in which data expressed in a certain language is not easily accessible to speakers of other languages.

The Web of Data shows the potential for being extended to a truly multilingual web as vocabularies and data can be published in a language-independent fashion, while associated language-dependent (linguistic) information supporting the access across languages can be stored separately. In this sense, the multilingual Web of Data can be realized in our view as a layer of services and resources on top of the existing Linked Data infrastructure adding i) linguistic information for data and vocabularies in different languages, ii) mappings between data with labels in different languages, and iii) services to dynamically access and traverse Linked Data across different languages.

In this article we present this vision of a multilingual Web of Data. We discuss challenges that need to be addressed to make this vision come true and discuss the role that techniques such as ontology localization, ontology mapping, and cross-lingual ontology-based information access and presentation will play in achieving this. Further, we propose an initial architecture and describe a roadmap that can provide a basis for the implementation of this vision.

Keywords: Multilingualism, Web of Data, Linked Data

1. Introduction

Tim Berners-Lee envisioned the Semantic Web as *"an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation"* [2]. Since then, the Web has witnessed an enormous growth in the amount of semantic information published in recent years, which has been stimulated to a large extent by the emergence of the Linked Data initiative [3,4]. Linked Data is a term referring to the recommended best practices for exposing, sharing, and connecting RDF [23] data via de-referenceable URIs on the Semantic Web. The crucial idea behind Linked data is to "connect data" using Semantic Web techniques and building on current Web infrastructure, thus transforming the Web into a "global database" in which resources are linked across sites and

where facts and related knowledge are available for consumption by advanced, knowledge-based web applications. In the rest of this paper we will use the term "Web of Data" exactly in this way. Linked Data has found a wide acceptance among governments[1], media companies, and academia all over the world [20,3,25]. These early adopters have clearly identified the potential benefits of publishing data in Linked Data format and are publishing their data sources following the Linked Data principles. The so called Linking Open Data (LOD) initiative[2] has further stimulated the emergence and adoption of the Linked Data principles in the construction of a web of linked open data.

Now that massive amounts of semantic data are becoming available on the Web, the question emerges how end users should access and interact with this wealth of data. As language is the most important means by which humans communicate, it is reasonable to assume that users would find a language-mediated way of accessing the Web of Data intuitive, appealing and effortless. In fact, the traditional Web is language-specific and information can only be accessed across languages if web sites are translated into the corresponding languages. In contrast, the Semantic Web can be assumed to be inherently language-independent, which means that information is given well-defined meaning by formally defining vocabularies or ontologies, building on semantic web languages such as OWL [1] or RDF [23]. In this sense, we argue that the Web of Data and hence the Semantic Web offer a great opportunity to make web information broadly accessible, independent of culture and native language. The main challenge involved in building this "Multilingual Semantic Web" is, however, to bridge the gap between language-specific information needs of users and language-independent semantic content.

In spite of the fact that much of the information available as linked data is by its nature language-independent, many resources also include linguistic information in the form of RDF(S) or SKOS [26] labels (*rdfs:label, skos:prefLabel*, etc.) which documents how a resource is named in multiple languages. Furthermore, by the use of standards such as the IETF language tags (RFC 5646), which builds on the ISO standard for representing language, script and region (ISO-639, 15924 and 3166 respectively), a general standard for the representation of labels in different languages, dialects, and writing systems has already been established. We also note here that some resources have linguistic information encoded within the label itself (generally, within the URI's fragment identifier). Furthermore, internationalized resource identifiers (IRI) were introduced by RFC 3987 to allow for non-ASCII characters to be included within these identifiers. However, although IRIs allow for constructing more human readable identifiers, they are not intended to represent complex linguistic information. We refer to [28] for an in-depth discussion on the use of local names and labels to describe resources in several natural languages.

Nowadays, most of the ontologies used in the Semantic Web as well as most of the data published according to the Linked Data principles have labels expressed in English[3] only, leading to an English bias in the Semantic Web which does not mirror the actual language distribution in the Web. It is likely that this situation will change as soon as more language communities start publishing their data in the Linked Data format. Indeed, our hypothesis is that this will be a growing trend shortly. At present, there are already many non-English Linked Data sources, e.g., Dogmazic[4], a French vocabulary of Creative Commons music,

[1] See for example http://ec.europa.eu/information_society/policy/psi/index_en.htm for a European initiative or http://www.data.gov/ for a US based one.

[2] http://esw.w3.org/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[3] E.g., Around 80% of ontology literals (with declared language) indexed in Watson are in English. See http://watson.kmi.open.ac.uk/blog/2007/11/20/1195580640000.html

[4] http://www.dogmazic.net/

GeoLinkedData.es[5], an open initiative to publish Spanish geospatial data in Linked Data (in Spanish, Catalan, and Galician, among other languages), as well as many other governmental initiatives originating from European Commission directives with the goal of making data of public interest available in an open fashion. Therefore, once monolingual resources in other languages are transformed into the Linked Data format, non-English monolingual Linked Data will increase in size over the years, creating "islands" of monolingual data. This will render the task of providing cross-language access to Linked Data even more challenging. These "islands" of monolingual Linked Data will result in situations in which access is restricted to speakers of the "right language". It is our standpoint that knowledge access which is restricted to speakers of certain languages does not match the open spirit of the Web of Data where most of the information is open and therefore accessible to anybody in principle. Thus, in this paper, we propose an infrastructure that has the potential of creating a level playing field offering equal opportunities for speakers of different languages when accessing web data.

In this article we pose ourselves several simple questions: How can universal access to the Web of Data be guaranteed to users, independent of the language they speak? How can the retrieval of Linked Data be supported if that data is distributed across unconnected monolingual data sources? With these questions in mind, one of the major challenges in future Semantic Web research will be to make sure that no matter in which language ontology terms are expressed, nor in which language the relevant data is published as Linked Data, cross-lingual access to knowledge and data has to be supported. Since the Web of Data is now becoming a reality, and the trend to transform databases into the Linked Data format is acquiring momentum, we need to be ready to manage multilingual access to ontologies and data. The main contribution of this paper is to identify and analyze such challenges as well as to propose a set of specific strategies in order to address them.

The remainder of the article is organized as follows. In Section 2, we present a motivating scenario and raise some specific research questions. Section 3 presents a proposal for architecture of services that might enable us to make the vision of a multilingual Web of Data true. A discussion on the different challenges involved in the creation of a multilingual Web of Data is presented in Section 4 and a roadmap of the tasks ahead is proposed in Section 5. Finally, conclusions are presented in Section 6.

2. Motivating Scenario

The Web has already transformed the way in which societies communicate, work, interact or even spend free time. We believe that it is time for the Semantic Web community to tackle the challenge of turning the Web into a multilingual Web of Data. In the following we will illustrate by means of a simple, hypothetical - but in our view realistic - example the different challenges that have to be addressed to realize this vision.

Let us imagine a Spanish tourist without a good command of German who is travelling in Germany. He realizes that he forgot his medication at home. He only remembers its commercial name in Spain (*Beglan*, a medicine to treat asthma) but not its active ingredients. He would like to find out about pharmacies or medical centres near his hotel in which the drug is available and which remain open until late. A potential query the user would have in mind could be: "*dame farmacias cercanas abiertas por la tarde y que dispongan de Beglan o alguna medicina equivalente*" (in English: "*give me nearby chemists that are open in the evening and have Beglan or an equivalent medicine in stock*").

The main difficulty in answering the query is the fact that: all information about *Beglan* will be available on Spanish sites, which will not contain any information about where the Spanish

[5] http://geo.linkeddata.es/

tourist will be able to find an equivalent drug in Germany close to his hotel. Secondly, German sites will contain the information about opening hours and available stock of pharmacies close to his hotel, but no information about *Beglan*. A suitable answer to this query can be found if either i) the German equivalent of *Beglan* is known somewhere on the Web (e.g. as a mapping between the URIs representing the two drugs) or ii) the German equivalents of the active ingredients contained in *Beglan* are known, such that a medicine with the same substances can be found.

In the multilingual Semantic Web that we envision, our tourist would query the Web of Data in his/her own language, i.e. Spanish, and would get the relevant data in that language due to the fact that the Linked Data cloud would be equipped with mechanisms for cross-language querying as well as presentation of results in any language. In particular, in our vision of the Multilingual Semantic Web, such mappings would be part of the Web of Data, linking islands of monolingual Linked Data together. Notice that the user in our example does not need to be aware of these semantic mechanisms acting behind the scenes.

This scenario raises several specific research questions related to the wider use of Linked Data. One area concerns the need to bridge the gap between the user's information needs and Linked Data from a multilingual perspective:

1. *How can a user pose questions in his/her own language to be processed against the web of Linked Data?* In our example, Spanish keywords such as "medicina" (Spanish word for *medicine*) or "farmacias" (*chemists*) have to be mapped to appropriate vocabulary elements independently of the language of the query, such that the intended meaning of the query can be formalized and transformed into an adequate SPARQL query, for instance. Research on multilingual question answering and multilingual semantic query construction will help to accomplish precisely this.

2. *How can Linked Data be delivered to ontology-based applications localized to different languages?* In particular, given a semantic query expressed in some ontological vocabulary, how can we find relevant resources from the Linked Data cloud that might not be expressed using the same vocabulary? In the example, imagine we deal with a semantic query on a Spanish vocabulary which asks for a list of medicines. Relevant data in German might be available but not directly linked to that Spanish ontology. Techniques to discover such data and establish cross-lingual links are needed to deliver data from sources in different languages.

3. *How should the results of a semantic query be adapted to the linguistic and cultural background of a user?* In our example, German data has to be translated back and verbalized into Spanish. More complex situations are also possible, such as languages using different literal representation (e.g., Chinese vs. German) or displaying rules (i.e., right-left vs. left-right). This will require the adaptation and localization of user interfaces and presentation views to a specific linguistic and cultural context.

The above mentioned questions are mainly concerned with user interaction. In order to support such an interaction which crosses language boundaries, a Linked Data web where resources are linked across languages is required as a basis. Important questions which arise in this context are the following:

4. *How can we adapt and translate the lexical/terminological layer of an existent ontology into other languages?* In order to expand the multilingual coverage of existing ontologies (to allow traversing and querying them no matter what language they are in), ontology translation and localization techniques have to be developed.

5. *Which approaches are suited to discover correspondences between ontologies expressed in different languages? Similarly, how can we align Linked Data across different natural languages?* This will involve ontology alignment techniques extended to deal with multiple languages (also required for research question 2). Let

us imagine that in our example we do not find a direct equivalence of *Beglan* with other medicines in Germany, but we discover its properties in a certain semantic dataset (e.g., its active ingredient is *salmeterol,* it is delivered as an inhalator, etc.). Based on the comparison of its properties, semantic mappings can be established with other datasets in German, thus supporting the discovery of the fact that *Aeromax* and *Serevent* are the German equivalents of *Beglan* because they share the same properties (active ingredients, via of administration, etc.).

6. *How to represent multilingual Linked Data?* We understand multilingual Linked Data as a set of resources in the Web of Data with associated linguistic information in several languages. Methods to represent and store multilingual Linked Data have to be devised. Current semantic standards offer ways of supporting this (e.g., *rdfs:label*) which might be enough for a great number of use cases. Nevertheless, richer models will be needed for supporting more demanding applications (e.g., information extraction from multilingual sources, verbalization methods, etc.) which can benefit from additional linguistic information such as lexical variation, part-of-speech, corpus provenance, etc.

7. *How to generate multilingual Linked Data?* There is ongoing work on the creation of Linked Data from relational databases and other forms of "data silos". Such methods have to be enriched to enable the generation of cross-lingual links, as well as the storage of multilingual lexical information in the representation models mentioned above.

The solutions to these research questions will contribute to realizing a Web of Data with advanced multilingual capabilities. In the remainder of the paper we analyse such open issues and identify a set of specific techniques that support the creation of a multilingual Web of Data.

3. A Web of Multilingual Linked Data

In Section 2 we have identified the principal issues that arise in dealing with multiple languages in the context of Linked Data, formulated in terms of research questions. In this section we reflect on these further and propose an architecture that has the potential to support the scenarios that we envision.

Although the Web of Data inherently supports multiple languages in the sense that it can represent any kind of data independent of any specific natural language, it still lacks explicit mechanisms to automatically exploit and reconcile such data in order to support access in any natural language. We understand the *multilingual Web of Data* as the current cloud of Linked Data enriched with a layer of services and resources that support and enhance the multilingual capabilities of the Web of Data with: i) linguistic information for different natural languages to be used in rendering the information contained in Linked Data sources, ii) mappings across Linked Data regardless of the language in which the data has been originally expressed, and iii) services to dynamically access and traverse Linked Data across languages.

In our view, the multilingual Web of Data will not be a new piece of Web infrastructure but will consist of services and models built on top of the current Web and Linked Data infrastructure. A possible architecture that instantiates this view will be discussed in the following paragraphs.

First of all, we identify some generic requirements for this architecture:

- The core of the architecture is the Linked Data cloud. This means in particular that publishing of and access to multilingual information on the Web of Data has to follow the Linked Data principles.

- The data remains untouched in the original Linked Data sources, while a further layer is added to account for linguistic, i.e., multilingual specification.
- The architecture will be arranged in two dimensions: i) multilingual *information* and ii) multilingual *services,* as will be explained in more detail below.

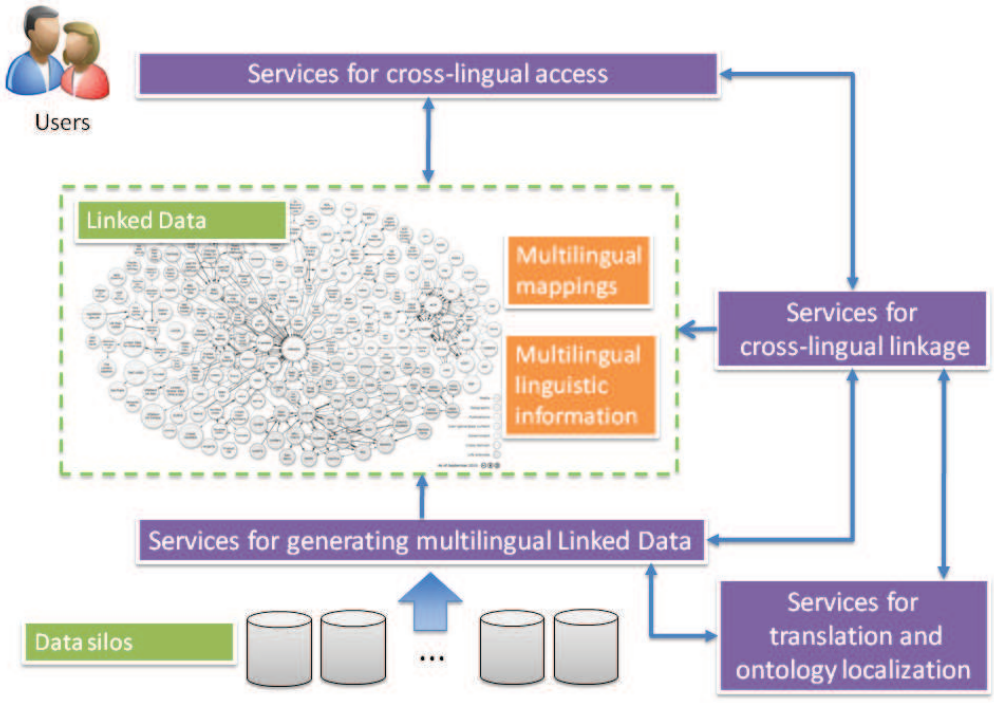


Figure 1: Architecture of a multilingual Web of Data, composed of multilingual and cross-lingual services and multilingual information around the cloud of Linked Data.

Figure 1 gives an overview of a possible architecture that fulfils the above requirements. The cloud of Linked Data is represented in the centre of the figure. The cloud is enriched with a layer of multilingual information (multilingual mappings and multilingual linguistic information) and with a set of services for creating, representing and accessing that multilingual information. On the bottom of the figure we represent a set of data silos that could be exported as Linked Data. These data silos can be monolingual or multilingual. Services for generating multilingual linked data are needed to transform these data silos into multilingual Linked Data. In addition, services are needed for the cross-lingual discovery and representation of mappings between Linked Data vocabularies and datasets expressed in different natural languages, as depicted on the right-hand side of the figure. Such cross-lingual linkage services are supported by localization services which translate or localize the vocabularies or ontologies used in the Linked Data cloud into several natural languages. A set of models to represent information about linguistic realization of vocabulary elements in different languages will also be needed to support more complex multilingual natural language processing applications. On top of the figure we represent the layer of cross-lingual access services that give the user access to the multilingual Web of Data. The following paragraphs explore this in some more detail.

3.1. Multilingual Information

In our proposed architecture, a multilingual Web of Data has to be supported by:

a. *Multilingual linguistic information* for generating and/or interpreting the multilingual realization of semantic definitions in Web ontologies and Linked Data. This includes

labels in multiple languages and can be enriched with additional linguistic information. For instance, imagine that *Beglan* is defined in an ontology; the simplest way of representing its lexical information in English might be stating that *onto:Beglan rdfs:label "Beglan"@es*. Richer lexical representations would allow indicating that "Beglan" is a proper noun, thus favouring smarter translation and verbalization strategies. This is of course beyond the expressivity of *rdfs:label* and will require additional modelling methods.

b. *Multilingual mappings* between Web ontologies/vocabularies that establish cross-lingual connections, as well as between linked datasets that use different languages in their lexical representation. Notice that these cross-lingual links can have associated linguistic information in several languages (so, in the most general case, we call them multilingual mappings). In particular, this information layer can contain semantic relationships between terms from ontologies in different languages (e.g., stating that onto1:Beglan is equivalent to onto2:Aeromax) or translations between lexical entries of ontology terms (e.g., "Medikament"@de is the German translation for "medicina"@es).

*Possible ways of rep*resenting this multilingual information are further discussed in Section 4.

3.2. Multilingual Services

The following set of services will support the creation of a multilingual network of linked data and its exploitation in order to provide access across languages:

a. *Services for generating multilingual linked data*. Publishing services are required to guide and support the creation of multilingual Linked Data on the basis of monolingual or multilingual databases. These services require cross-lingual linking services to operate and produce as output new Linked Data enriched with multilingual mappings and associated linguistic information. Typically, these processes will be run off-line and might involve user supervision. Generation services can also benefit from ontology localization techniques that will be used to create multilingual vocabularies and datasets from monolingual ones.

b. *Services for translation and ontology localization.* They will allow an automatic translation of the vocabularies and ontologies used in the Linked Data cloud. These services will have to rely on models to represent multilingual linguistic information associated to vocabularies and ontologies (and potentially also to datasets) in the cloud. Localization services can serve as means for supporting the cross-lingual linking of vocabularies and datasets in the Linked Data cloud.

c. *Services for cross-lingual linkage*. A family of services has to be developed for discovering cross-lingual mappings (i.e., relationships between ontology terms or semantic data expressed in different languages) and multilingual mappings (i.e., cross-lingual mappings that can be multilingual themselves). This includes the development of ontology mapping techniques that can align vocabularies in different languages as well as techniques to link datasets in different languages. These services will allow for cross-lingual mapping to be discovered off-line. However, we also foresee the case in which translation and localization services are dynamically accessed, and cross-lingual mappings are computed on the fly, pre-populating the required multilingual infrastructure (multilingual mappings and linguistic annotations) to be associated with Linked Data (similarly to the periodical updating of DBpedia).

d. *Services for cross-lingual access*, supporting the access of the cloud of Linked Data independently of the natural language used. This family of services will include techniques for semantic query generation, question answering, query federation, visualization methods, etc. They will operate typically at run-time by dynamically processing and responding to cross-lingual queries on the Linked Data. As a side effect, this will also incrementally populate the multilingual infrastructure (multilingual mappings and linguistic annotations), which will be made persistent and stored for future reuse.

While we have emphasized the need for (semi-) automatic approaches, it is desirable to create strong incentives so that people feel motivated to participate in the endeavour of creating a multilingual Web of Data by manually defining mappings, publishing lexica for their favourite ontologies in their language, verifying the quality of translations, etc. Participation of users forms part of the success story of the Web. People in fact do not only contribute content to the Web in the form of websites, they also create links between pages, contribute questions and answers on blogs and forums, upload videos and images, etc. People are thus willing to contribute to the Web with their own resources. A big challenge for the multilingual Web of Data as sketched in this article is to create appropriate incentives for people to devote resources and to contribute to this endeavour as well.

Besides creating new multilingual resources compliant with the Linked Data principles, it seems obvious that already existing ones can also play a principal role. This is the case for EuroWordNet[6] or the various WordNets that have been created in different natural languages in subsequent projects (GlobalWordNet[7], Meaning[8], BalkaNet, etc.), and which have been linked to the so-called InterLingual Index, a set of common concepts initially obtained from the Princeton WordNet. Since WordNet is one of the resources that have been included in the Linked Data cloud from early on, it is not unreasonable to think that the rest of WordNets will follow the same transformation process. Such a multilingual resource could be used with two purposes: 1) as a multilingual dataset that can be accessed in different languages in the Linked Data cloud, and 2) as a multilingual lexical resource that can be reused within the Linked Data cloud to enrich available datasets with lexical information, or even to help in the process of establishing mappings between multilingual datasets in the cloud. However, WordNet contains mainly domain-independent concepts and does not cover lexical information on specific domains of knowledge. Hence, other models are required to directly enrich datasets in the cloud, unless specific WordNets containing domain lexical elements are created and linked to datasets in the cloud. Since domain specific WordNets are already available for some domains (see KYOTO project[9]), this approach could be adopted for Linked Data. The question would then be if the lexical-semantic information captured in WordNets would suffice, or if further models of linguistic and terminological descriptions should be devised to enrich the semantics captured in the Linked Data cloud.

4. Challenges

In order to make the vision of multilingual Linked Data come true, the specific research questions introduced in Section 2 have to be addressed. These questions provide a natural roadmap for the development of a multilingual Web of Data. The issues that these questions

[6] http://www.illc.uva.nl/EuroWordNet/

[7] http://www.globalwordnet.org/

[8] http://adimen.si.ehu.es/wei4/doc/mcr/meaning.html

[9] http://www.kyoto-project.eu/

raise rely upon four core challenges, namely: ontology localization, multilingual mappings, representation of multilingual Linked Data, and cross-lingual access to the Web of Data.

4.1. Ontology Localization

Inspired by software localization [12], ontology localization [10,13,36] has been defined as the process of adapting an ontology to the needs of a particular (linguistic and cultural) community. In this context, some techniques have been developed with the main goal of translating the terminological layer of the ontology, so that it can be reused in multiple linguistic (and cultural) scenarios (see [14,15]). A *localized* ontology can be understood as an ontology adapted to the target community and language, that is used independently from the original ontology. In this case, localization would be understood in line with the localization of software products, in which one or more monolingual ontologies are obtained as end product. However, it can also be understood as an ontology in which its terminological layer has been translated to one or several target languages, and in which the different linguistic versions interoperate, resulting in a multilingual ontology. As has been identified in [10,13], the translation of the terminological description of ontologies may not always suffice to make the ontology reusable in different linguistic and cultural scenarios. This fact is mainly related to the domain represented by the ontology, whose conceptualization may not always be compliant with the needs of the target community. In this case, modifications or adaptations of the conceptualization would be needed, although this will not be further dealt with in this paper. Some ontology localization tools already exist, such as LabelTranslator [14,15]. In its current version, this system uses a translation service that obtains automatic translations for each ontology label by consulting a number of linguistic resources, such as multilingual lexica (EuroWordNet), multilingual terminologies (IATE[10], a multilingual term base of the EU), and translation services (Babelfish[11], GoogleTranslate[12]). Then, candidate translations are disambiguated according to the ontological context. The translations obtained for the different ontology elements are then stored in an external model that associates a set of linguistic descriptions in several natural languages to ontology elements, resulting in a multilingual ontology. The Protégé plugin OntoLing [30,31,32] instead provides an environment for manual enrichment of ontology classes with linguistic information stored in a set of dictionaries that can be accessed from the plug-in API.

In order to support the full automatic localization of vocabularies and ontologies, more research is needed in the area of ontology localization so that the accuracy of current approaches can be increased. Further, techniques will have to reduce the cost of localizing in order to be usable at Web scale. Furthermore, localization tools have to be user-friendly and integrated into standard work practices[13].

4.2. Cross-lingual Mappings

There are three levels at which relationships between ontology terms or semantic data with labels in different languages can be established:

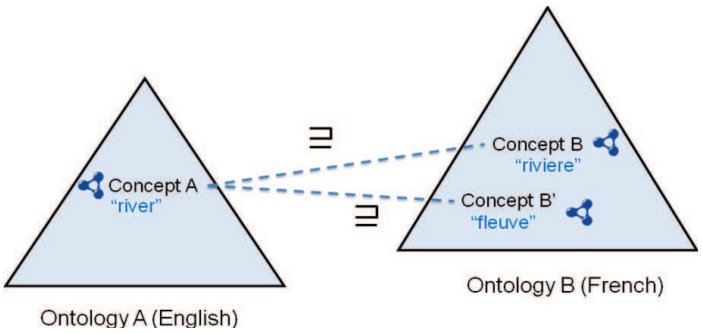
[10] http://iate.europa.eu

[11] http://babelfish.yahoo.com/

[12] http://translate.google.com/

[13] The EU funded project Monnet on Multilingual Ontologies for Networked Knowledge has a focus on such development: http://www.monnet-project.eu/

1. *Conceptual level*. At this level, concepts from different ontologies described in different languages can be semantically related by using ontology constructs, either to represent taxonomical relations (i.e., *owl:equivalentClass, owl:sameAs, rdfs:subClassOf,* etc.) or domain dependent relations (i.e., ontology properties coming from other ontologies). We call them *conceptual cross-lingual mappings*. Such links permit to establish a correspondence between or among concepts included in different ontologies, and which are described in the same or in a different language. Consider for example an ontology of the hydrographical domain and in particular the relation between the concept "watercourse" in an ontology documented in English and "cours d'eau" in an ontology documented in French. Both labels are referring to the same concept, but they are expressed in different natural languages. We could also make use of the *rdfs:subClassOf* relation if an ontology documented in a different or the same language would contain a world phenomenon described with a higher granularity level. Let us take the case of the concepts "fleuve" and "rivière" in the same ontology of the hydrographical domain in French. "Fleuves" are rivers that flow into the sea, whereas "rivières" can be defined as rivers that flow into the sea or into another stream. Both lexicalized concepts in French do not have an exact equivalence or direct correspondence in English, but its closest concept is described by "river" in English, which subsumes both concepts. This has been illustrated in Figure 1.
2. *Instance level*. At this level, links are established between individuals instead of between their associated concepts. Thus, we call them *instance cross-lingual mappings*. At this level also ontological constructs such as *owl:sameAs* can be used to represents the cross-lingual mappings. For example we can state that "Spain" in the English dataset GeoNames[14] is the same as "España" in the Spanish dataset GeoLinkedData[15]. Other types of relations can be used instead when appropriate, for example to state that *geolinkeddata:Madrid isCapitalOf geonames:Spain*. This relation can be multilingual itself, thus corresponding to the notion of multilingual mapping introduced above. For instance the property *isCapitalOf* can be associated to labels *"is capital of"@en* and *"es capital de"@es*.
3. *Linguistic level*. Here the links would not be established between the concepts (or instances) themselves but between their associated linguistic information. We call them *linguistic cross-lingual mappings*. This sort of mappings can be very useful when keeping uncoupled the conceptual and linguistic information is a major requirement. In order to allow two ontologies to interoperate at the linguistic level, mappings would be established between the linguistic descriptions of their concepts, which are not necessarily exact equivalents but the closest correspondences between culture-specific concepts. In this case, no semantic relation is established between the concepts as in the cases described above. In the simplest case, a property labelled "translation" or "cultural equivalent" (for instance) might be established between the lexical realizations of the concepts.



[14] http://sws.geonames.org/2510769/

[15] http://geo.linkeddata.es/page/resource/Pa%C3%ADs/Espa%C3%B1a

Figure 1. Example of cross-lingual conceptual mappings.

Figure 2 shows a possible solution for level 3 in which the linguistic relations are mediated by external lexical information. We call *senses* the relations established between concepts and their lexicalizations. This link restricts the meaning of lexical entries (e.g., "river") in the specific context of the ontology, thus providing disambiguation among the potential senses. In this way, concepts in the ontology point to lexical entries in monolingual lexicons. Translations can be inferred between lexical entries in the different monolingual lexicons. The main advantage of such an approach is that the relations between labels in different languages can be made explicit at the sense level, which would be in charge of dealing with inter-cultural and inter-linguistic disparities.

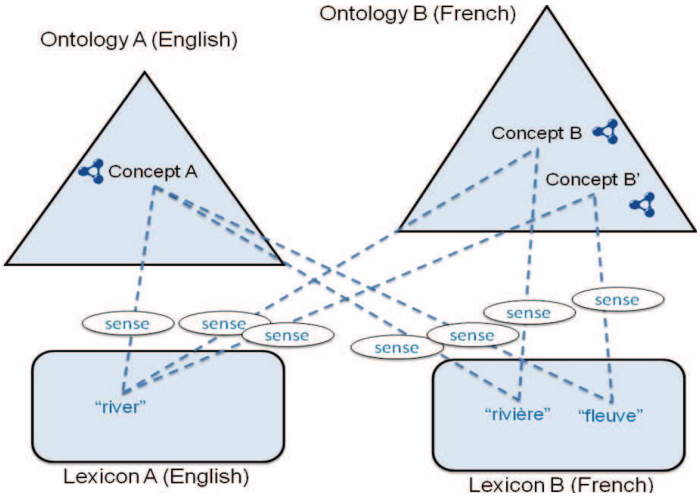


Figure 2. Example of cross-lingual linguistic mappings mediated by lexical information

Regarding the way cross-lingual ontology mappings (levels 1 and 2) can be represented, we consider current Semantic Web languages and ontology matching formats rich enough to support this. Indeed, cross-lingual ontology mappings can be considered a sub-case of ontology mappings. On the other hand, representation of cross-lingual linguistic mappings (level 3) would need a more careful study involving ontologies describing translations[16] and definition of online lexica based on Semantic Web principles [24].

Besides representing cross-lingual mappings, another major issue is how to discover them. While many approaches to the automatic identification of mappings are available (so called ontology alignment or matching [16]), there are not many approaches that identify such mappings between ontologies with labels in different languages. Ontology matching approaches have to be extended to be able to discover matches between vocabularies in different languages. It seems plausible to assume that current techniques can be extended to support multiple languages without requiring the development of radically different approaches. In fact, most of the extant current ontology matching systems rely to some extent on comparisons based on lexical information and could be extended by allowing for cross-language comparisons by either integrating machine translation systems (compare [17,39,35]) or multilingual lexical resources. Other matching systems rely on the computation of semantic similarity or relatedness between entities. Thus, cross-lingual measures of semantic relatedness or similarity need to be developed to support multilingual ontology matching and

[16] E.g., LOD ontology http://semlabs.upmc.fr/LODInTranslation/LodInTranslationOntology.owl

cross-lingual link discovering. In [11,27] some existing monolingual techniques are adapted for their use with bilingual lexica as a source of background knowledge, showing promising results. A broader approach [19] has been proposed for computing cross-lingual relatedness using the inter-language links contained in Wikipedia. The potential use of existing monolingual measures when applied to multilingual contexts has to be further analysed and new cross-lingual semantic measures have to be devised. A very interesting research avenue is the development of semantic similarity and relatedness measures that - in contrast to the ones described in [11,27,19] - do not rely on particular lexica or knowledge sources but exploit the Linked Open Data cloud or the Web as a whole as corpus.

Eventually, multilingual mappings should be shared on the Web for exploitation by cross-lingual information access methods and other techniques that require traversing multilingual ontological information. Although several solutions are possible here, a natural way of doing this is to represent and store such mappings in a way that is compliant with the Linked Data principles, thus becoming part of the Linked Data cloud itself. In principle, owing to the distributed nature of the Semantic Web, anyone could upload their discovered mappings to the cloud.

4.3. Representation of multilingual lexical information

Vocabulary elements can be realized in a variety of different ways across languages. The *rdfs:label* property can capture such variance at the term level, but is not sufficient to capture syntactic variation across languages. For example, a certain property "border" might be represented as a transitive verb in English ("to border"), via a verb with a prepositional phrase in German ("grenzt an") and as a construction involving the verb "to have" in Spanish ("tiene frontera con"). Such different constructions by which the same property can be realized across languages cannot be represented using the *rdfs:label* property. Further, linguistic cross-lingual mappings cannot be directly expressed in RDF(S) as they would require the reification of labels in order to say that "river" is a "translation equivalent" of "riviere" and "fleuve". Models which feature the required expressivity to represent such multilingual lexical information are needed. While such models exist already (e.g., LexInfo [8,9,7] or LIR [33,29]), they need to be modularized so that those applications in need of only simple models will not be forced to deal with the full complexity of these models. The lexicon model for ontologies (*lemon* [24]) is organized already in such a modular fashion and thus will be usable by applications requiring different degrees of expressivity. Tool support for these models as well as more convincing applications exploiting these models will be needed in the near future to convince stakeholders in the Web of Data to adopt such models.

4.4. Cross-lingual access to and querying of Linked Data

Currently, the standard way for accessing RDF data is by means of SPARQL [34]. Appropriate mechanisms will be needed to facilitate querying of multilingual data in the LOD building on SPARQL. In particular, available mappings between vocabulary elements in different languages should be taken into account in order to, given a query in a particular language, retrieving answers of the same, some other or even multiple languages. Further, the extension of SPARQL endpoints with descriptions specifying the supported languages will facilitate tasks such as query federation [6], planning, and source selection in a multi-lingual context.

Nevertheless, the use of SPARQL is not intended for end-users, who are more used to other ways of interaction, such as keyword-based queries. How to transform a set of user keywords or a query posed in natural language into a formal query is still a research question that has attracted the attention of many recent research efforts [21,37,38,5,22]. These techniques have been initially conceived to operate with a single language, and some of them prefer to use

monolingual English resources such as WordNet. Nevertheless, we do not find theoretical limitations in current question answering and semantic query construction systems that prevent its use across languages. Some first steps to be explored towards the realization of this goal are the use of cross-lingual semantic measures and the exploitation of multilingual background knowledge sources.

A further option that we envision is to allow ontology-centred applications to access that subset of Linked Data that is relevant to the ontology in question, even if the languages of the ontology and the data do not match. This will require; i) the discovery of multilingual mappings between a given ontology in one language and the ontologies used in Linked Data, typically with descriptions in English[17] or even other languages in the future; ii) the creation of consistent views over Linked Data with results relevant to that ontology. Another interesting scenario arises when instead of relying on one predefined ontology, several ontologies are dynamically accessed to describe the semantics of a user query [38]. In this case we have to deal with the problem of redundancy (different ontological descriptions representing the same meaning) [18]. Techniques to cope with this issue are thus required, especially in a multilingual environment, where we need to capture for instance that ontologies referring to the Spanish term "medicina" are describing the same concept as ontologies defining the English equivalent "medicine".

Finally, content negotiation, used in the Web for offering different formats and language versions of the same Web document, can play a role when localised semantic information has to be served to the final user or application.

5. Roadmap

A roadmap towards making the multilingual web of data come true could look as follows:

Stage I. Lightweight models for representing multilingual lexical information will be available in combination with a first generation of ontology localization services that support the translation of the labels of some vocabulary or ontology into another language. Simple techniques for inferring links across vocabulary with labels in different languages will be available. These first techniques might make use of multilingual lexical (EuroWordNet) and terminological resources as well as extant translation services to discover such mappings. Early applications using the above mentioned multilingual knowledge (e.g. to automatically localize semantic web pages) or providing cross-lingual query support will provide the required incentives for the development of more complex infrastructure in future stages. Furthermore, current existent techniques to migrate data silos into the Linked Data format will be extended to enrich the produced semantic data with multilingual mappings.

Stage II. Semantic search engines might index multilingual lexical information available on the Web as well as integrate available ontology localization tools to support answering ad-hoc queries in any language. More complex models for representing multilingual lexical correspondences will be available, supporting cross-lingual natural language processing applications requiring deeper multilingual lexical knowledge.

Stage III. Users might recognize the benefit of the provision of multilingual lexical information offers, a situation that might create the appropriate incentives to motivate people to add and evaluate manual and (semi)automatic mappings between vocabularies in different languages. Further, an ecosystem of services that accomplishes various tasks will be available as part of the Web of Data. Services that can be exploited for cross-language querying will provide on-demand translation to be exploited by an advanced set of techniques for

[17] As can be confirmed by taking a look at recommended vocabularies in Linked Data (http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/#whichvocabs)

establishing cross-lingual mappings. Building on more complex models for representing multilingual linguistic information, search engines might be able to process natural language questions in any language. Semantic search engines will be able to adapt their result presentation to conventions of the linguistic and cultural community to which the user belongs.

In working towards this roadmap, several research fields should be involved in order to benefit from their expertise. Especially interesting is the expertise available in the database community with respect to the representation of multilingual information in databases. The computational linguistics and lexical resources community could contribute their models for the representation of multilingual lexicons such as LMF. From the area of machine translation, different approaches could be reused to develop systems that are able to translate the lexical layer of ontologies. Techniques from word sense disambiguation as developed in the natural language processing community will also play a major role here as identification of the right senses of a label in a given ontology might allow for more accurate translations.

6. Conclusions

The traditional Web is characterized by the fact that information is only available across languages if web sites are translated, thus necessarily leading to duplication of information. The Web of Data has the potential for extending the Web to a truly multilingual Web as information can be published in a language-independent fashion using the RDF data model and following the Linked Data principles. The multilingual Web of Data can be realized, in our view, as a layer of services and resources on top of Linked Data which adds i) linguistic information in different languages ii) mappings between data with labels in different languages, and iii) services to dynamically access and traverse the Linked Data across different languages. All these extensions are challenging, but current Semantic Web and Linked Data technologies are mature enough to render the vision feasible.

In this paper we have discussed the challenges involved in building a multilingual Web of Data and proposed a general architecture for realizing this. A multilingual Web of Data seems feasible given the current state of Semantic Web Technology, but in order to build it we need to understand that in the same way that much of the traditional Web is a Web for people made by people as a collective, a multilingual Web of Data can at best only provide the infrastructure which needs to be "populated" by people, adding mappings between vocabularies in different languages, lexica for ontologies etc. This is, indeed, one of the most challenging aspects of the roadmap towards a multilingual Web of Data envisioned in this paper.

Acknowledgements

We thank Dr. Ignacio García-Álvarez for his kind assistance with our motivating example. This work is supported in part by the European Union under Grant No. 248458 for the Monnet project as well as by the Centro Nacional de Información Geográfica and CDTI under the R&D programme Ingenio 2010 for the R&D project España Virtual, by the Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2), and by the CICYT project TIN2010-17550 named “Multilingüismo en Ontologías y Linked Data” funded by the Spanish Ministry of Science and Innovation.

References

1. S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D. L. Mcguinness, P. F. Patel-Schneider, and L. A. Stein. OWL Web Ontology Language Reference. Technical report, W3C Recommendation, February 2004.

2. T. Berners-Lee, J. Hendler and O. Lassila. The Semantic Web. Scientific American, May 2001.
3. C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. International Journal on Semantic Web and Information Systems (IJSWIS), 2009.
4. C. Bizer, T. Heath, K. Idehen, and T. Berners-Lee. Linked data on the web (LDOW2008). In proceedings of the 17th international conference on World Wide Web, pages 1265-1266, Beijing, China, ACM, 2008.
5. C. Bobed, R. Trillo, E. Mena and J. Bernad. Semantic Discovery of the User Intended Query in a Selectable Target Query Language. Proc. of 7th International Conference on Web Intelligence (WI 2008), Sydney (Australia), IEEE Computer Society Press, pp. 579-582, December 2008.
6. C. Buil, M. Arenas, and O. Corcho. Semantics and optimization of the SPARQL 1.1 federation extension. In Proc. of 8th Extended Semantic Web Conference (ESWC 2011), Heraklion, Crete, Greece, Lecture Notes in Computer Science, vol. 6644, pp. 1-15, Springer, 2011.
7. P. Buitelaar, P. Cimiano, P. Haase, M. Sintek, Towards Linguistically Grounded Ontologies. In: Proceedings of the 6th European Semantic Web Conference (ESWC 2009), 2009.
8. P. Cimiano, P. Buitelaar, J. McCrae, M. Sintek: LexInfo: A declarative model for the lexicon-ontology interface. J. Web Sem. 9(1): 29-51 (2011)
9. P. Cimiano, P. Haase, M. Herold, M. Mantel, and P. Buitelaar. Lexonto: A model for ontology lexicons for ontology-based nlp. In Proceedings of OntoLex'07, Busan, South Corea, 2007.
10. P. Cimiano, E. Montiel-Ponsoda, P. Buitelaar, M. Espinoza, A. Gómez-Pérez. A Note on Ontology Localization - Journal of Applied Ontology 5(2), 2010.
11. S. Eger and I. Sejane. “Computing Semantic Similarity from Bilingual Dictionaries.” In Proceedings of the 10th International Conference on the Statistical Analysis of Textual Data (JADT-2010), pages 1217-1225, Rome, Italy, June 2010. JADT-2010.
12. B. Esselink. A practical guide to localization. John Benjamins, 2000.
13. M. Espinoza, E. Montiel-Ponsoda and A. Gómez-Pérez. Ontology Localization. Proceedings of the 5th Fifth International Conference on Knowledge Capture (KCAP2009) in Redondo Beach, California, USA, pp. 33-40. 2009.
14. M. Espinoza, A. Gómez-Pérez, and E. Mena. Enriching an ontology with multilingual information. In Proc. of 5th European Semantic Web Conference (ESWC'08), Tenerife, (Spain) LNCS Springer, pp. 333-347. 2008.
15. M. Espinoza, A. Gómez-Pérez, and E. Montiel-Ponsoda. Multilingual and Localization Support for Ontologies. In Proceedings of 6th European Semantic Web Conference (ESWC’09), Heraklion (Grecia), Springer Verlag LNCS, pp. 821-825, demo paper, 2009.
16. J. Euzenat and P. Shvaiko. Ontology matching. Springer-Verlag, 2007.
17. B. Fu, R. Brennan, and D. O'Sullivan. Cross-lingual ontology mapping - an investigation of the impact of machine translation. in ASWC, ser. Lecture Notes in Computer Science, vol. 5926, pp. 1-15, Springer, 2009.
18. J. Gracia, M. d'Aquin, and E. Mena. Large Scale Integration of Senses for the Semantic Web. Proc. of 18th International World Wide Web Conference (WWW 2009), Madrid, Spain, ACM, pp. 611-620, April 2009.
19. S. Hassan and R. Mihalcea. “Cross-lingual Semantic Relatedness Using Encyclopedic Knowledge.” In Proceedings of the conference on Empirical Methods in Natural Language Processing, Singapore, 2009.
20. G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer, and R. Lee, Media meets semantic web - how the bbc uses dbpedia and Linked Data to make connections. In ESWC 2009 Heraklion: Proceedings of the 6th European Semantic Web Conference on The Semantic Web. Berlin, Heidelberg: Springer-Verlag, pp. 723-737, 2009.
21. Y. Lei, V. Uren and E. Motta. SemSearch: A Search Engine for the Semantic Web. EKAW, Springer, volume 4248, pp. 238-245, 2006.
22. V. López, M. Sabou, and E. Motta. PowerMap: Mapping the Real Semantic Web on the Fly. In Proc. of 5th International Semantic Web Conference, Athens, GA, USA, November 5-9, 2006, LNCS 4273, 2006.
23. F. Manola and E. Miller. RDF Primer. Technical report, W3C Recommendation, February 2004.
24. J. McCrae, G. Aguado-de-Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gómez-Pérez, J. Gracia, L. Hollink, E. Montiel-Ponsoda, D. Spohr, T. Wunner. Interchanging Lexical Resources in the Semantic Web. Language Resources and Evaluation, in press (2011).
25. P. Miller. Linked data horizon scan. Joint Information Systems Committee (JISC), Tech. Rep., December 2009
26. A. Miles, S. Bechhofer. SKOS-Simple Knowledge Organization System Reference, W3C Recommendation, August 2009.
27. S. Mohammad, I. Gurevych, G. Hirst, and T. Zesch. Cross-Lingual Distributional Profiles of Concepts for Measuring Semantic Distance. In Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 571-580. ACL, June 2007.
28. E. Montiel-Ponsoda, Daniel Vila-Suero, Boris Villazón-Terrazas, Gordon Dunsire, Elena Escolano and Asunción Gómez-Pérez. Style Guidelines for Naming and Labeling Ontologies in the Multilingual Web. Proceedings of the DCMI International Conference on Dublin Core and Metadata Applications (DC-2011), The Hague, The Netherlands. 2011.
29. E. Montiel-Ponsoda, G. Aguado de Cea, A. Gómez-Pérez, and W. Peters. Modelling multilinguality in ontologies. In Coling 2008: Companion volume - Posters and Demonstrations, pp. 67-70, Manchester, UK, 2008.

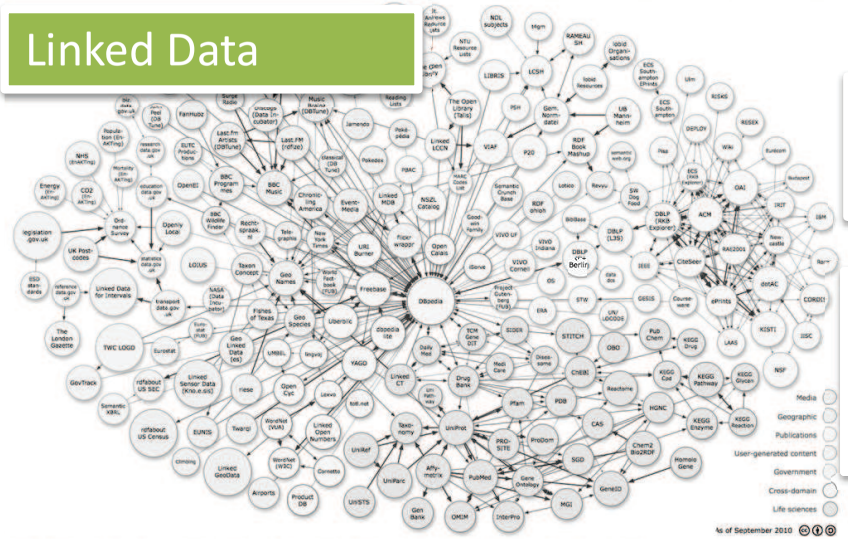
30. M.T. Pazienza and A. Stellato. The Protégé Ontoling Plugin - Linguistic Enrichment of Ontologies in the Semantic Web. Poster proceedings of the 4th International Semantic Web Conference (ISWC-2005) Galway, Ireland, 2005.
31. M.T. Pazienza and A. Stellato. Linguistic Enrichment Ontologies: a methodological framework. Second Workshop on Interfacing Ontologies and Lexical Resources for Semantic Web Technologies (OntoLex2006), held jointly with LREC2006, Magazzini del Cotone Conference Center, Genoa, 2006.
32. M.T. Pazienza and A. Stellato. An open and scalable framework for enriching ontologies with natural language content. The 19th International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA/AIE'06), special session on Ontology & Text Annecy, France, 2006.
33. W. Peters, E. Montiel-Ponsoda, and G. Aguado de Cea. Localizing Ontologies in OWL. In Proceedings of OntoLex'07, Busan, South Corea, 2007.
34. E. Prud'hommeaux and A. Seaborne. SPARQL query language for RDF. W3C Recommendation, Tech. Rep., January 2008
35. D. Spohr, L. Hollink and P. Cimiano. A machine learning approach to multilingual and cross-lingual ontology matching. In Proc. of ISWC 2011, to appear. Bonn, Germany, October 2011
36. M.C. Suárez-Figueroa and A. Gómez-Pérez. Towards a Glossary of Activities in the Ontology Engineering Field. Proceedings of the Sixth International Language Resources and Evaluation (LREC'08). Marrakech, May 2008.
37. T. Tran, P. Cimiano, S. Rudolph, and R. Studer. Ontology-Based Interpretation of Keywords for Semantic Search. Proc. of 6th International Semantic Web Conference (ISWC'07), Busan, Korea, Springer, volume 4825, pp. 523-536, 2007.
38. R. Trillo, J. Gracia, M. Espinoza, and E. Mena. Discovering the Semantics of User Keywords. Journal on Universal Computer Science (JUCS). Special Issue: Ontologies and their Applications, ISSN 0948-695X, 13(12):1908-1935, Springer Verlag, December 2007.
39. C. Trojahn, P. Quaresma, and R. Vieira. A framework for multilingual ontology mapping. In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08). Marrakech, Morocco: European Language Resources Association (ELRA), May 2008.

Figure

Users

Services for cross-lingual access

Linked Data



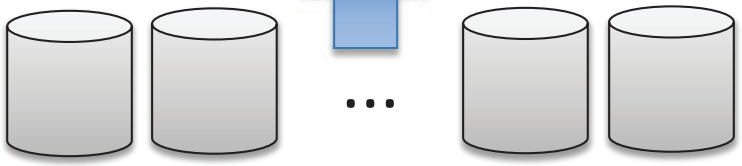
Multilingual mappings

Multilingual linguistic information

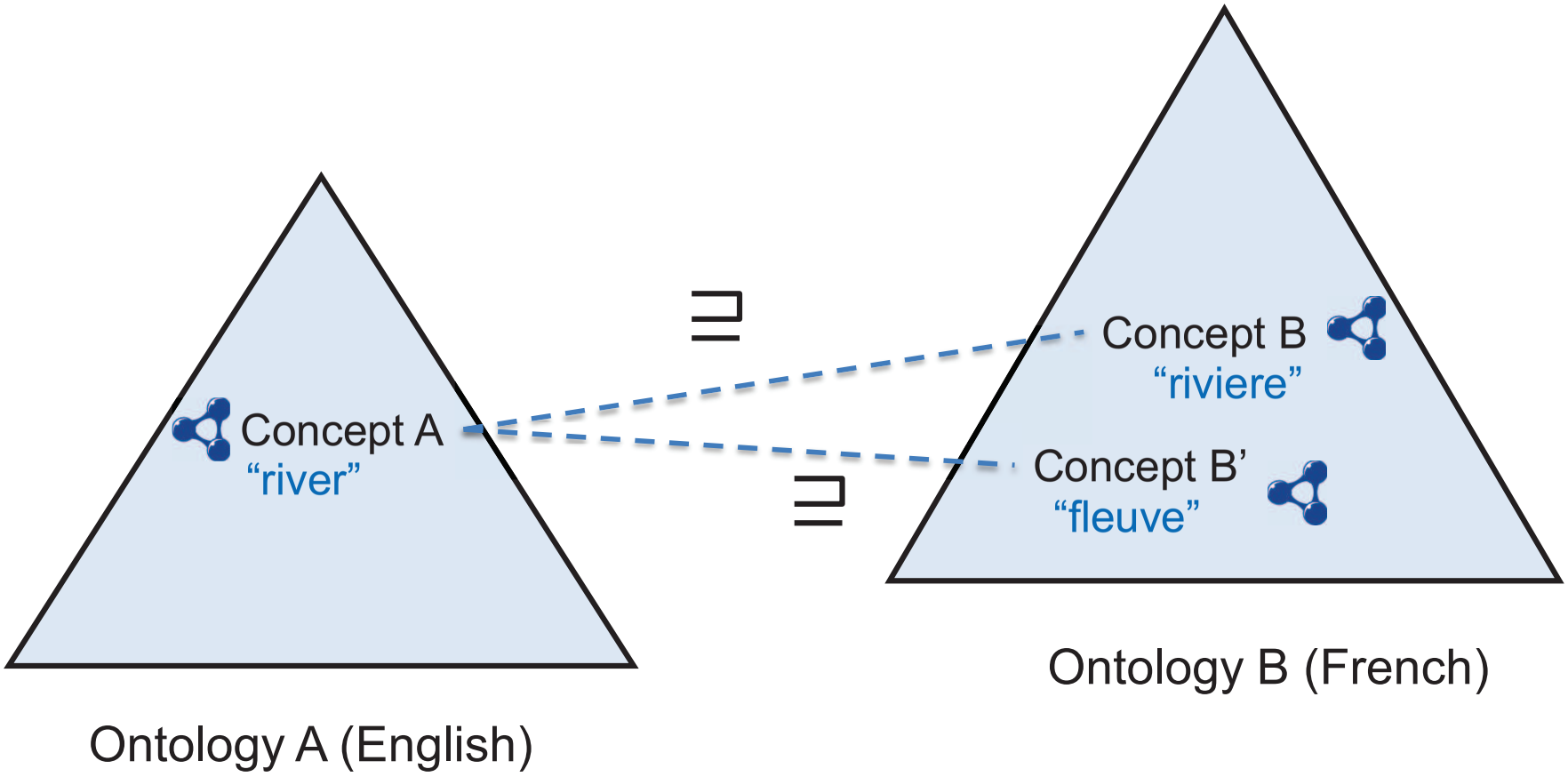
Services for cross-lingual linkage

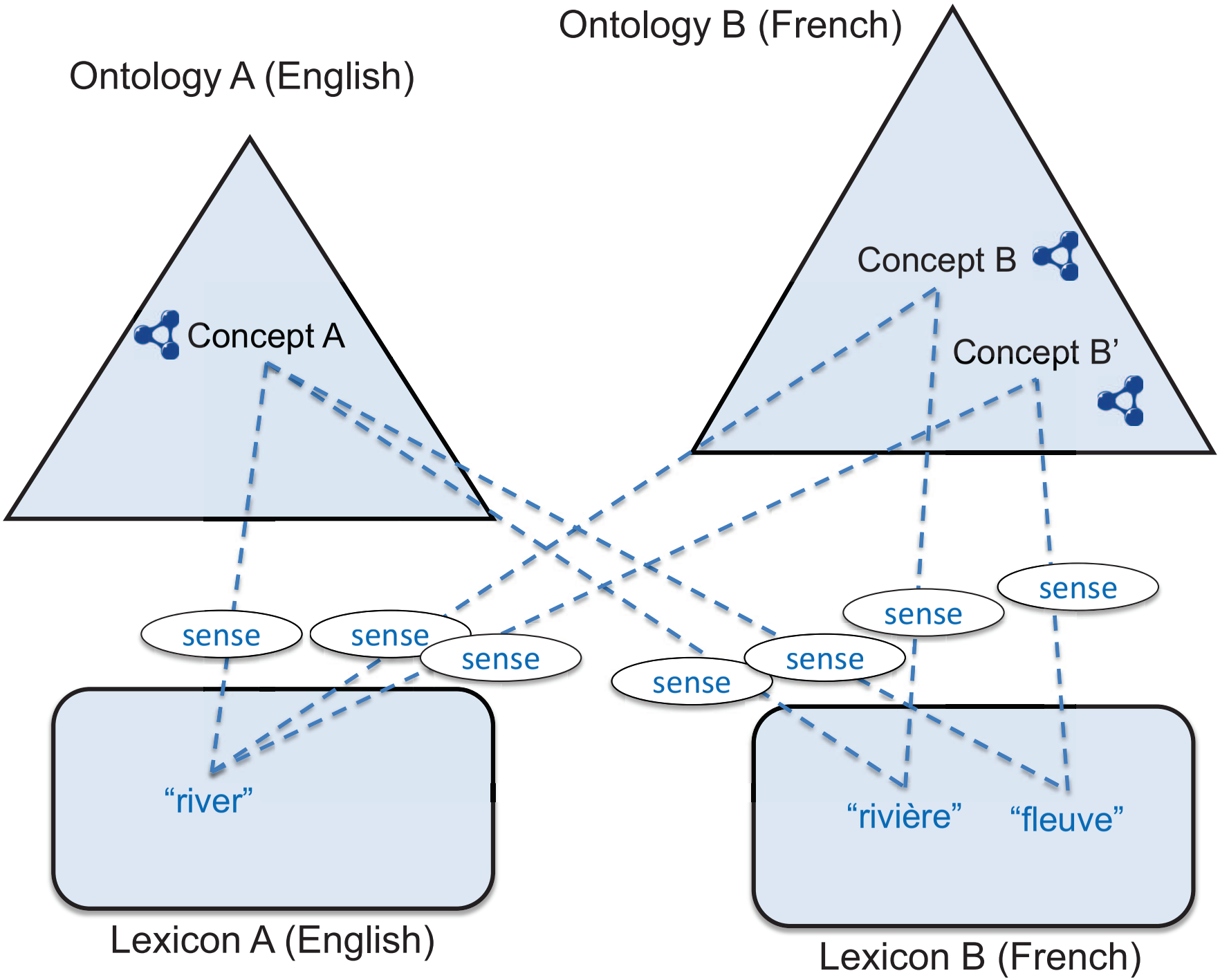
Services for generating multilingual Linked Data

Data silos



Services for translation and ontology localization


Concept A
“river”
⊇
⊇
Concept B
“riviere”
Concept B’
“fleuve”
Ontology A (English)
Ontology B (French)


Ontology A (English)
Ontology B (French)
Concept A
Concept B
Concept B'
sense
sense
sense
sense
sense
sense
"river"
"rivière"
"fleuve"
Lexicon A (English)
Lexicon B (French)