Ontology-based Multilingual Access to Financial Reports for Sharing Business Knowledge across Europe

Thierry Declerck & Hans-U. Krieger[1], Susan M. Thomas[2], Paul Buitelaar & Sean O'Riain & Tobias Wunner[3], Gilles Maguet[4], John McCrae & Dennis Spohr[5], Elena Montiel-Ponsoda[6]

[1]*DFKI GmbH, Stuhlsatzenhausweg, 3, D-66123 Saarbrücken, Germany*
[2]*SAP AG, Research, Vincenz-Priessnitz-Straße 1, D-76131 Karlsruhe, Germany*
[3]*DERI, National University of Ireland, Galway, Newcastle Rd, Galway, Ireland*
[4]*XBRL-Europe, Avenue d'Auderghem, 22-28/8, B-1040 Brussels, Belgium*
[5]*CITEC – University of Bielefeld, Universitätsstr. 21-23, 33615 D-Bielefeld, Germany*
[6] *UPM – Facultad de Informática, Boadilla del Monte, Madirid, Spain*

Abstract: Within the FP7 European project MONNET -- Multilingual Ontologies for Networked Knowledge see http://www.monnet-project.eu/ --, we are specifying and implementing a use case concerning business intelligence on European companies, involving a semantic-level analysis of business reporting in several languages. This use case is building on national and international accounting regulations that are encoded in XBRL taxonomies. XBRL (eXtensible Business Reporting Language) is an XML-based open standard for identifying and communicating complex financial information in corporate business reports.

In Monnet we plan to use an "upgraded" XBRL in form of localized ontologies that support not only the translation of the central elements of business reports (in Dutch, English, German and Spanish), but also the extraction, integration and presentation of financial data available in various types of documents in various languages. With this use case, Monnet hopes to contribute to the effective sharing of financial and business knowledge across Europe.

1. Introduction

A universal response to the current financial crisis was the call for more transparency on the part of banks and investment firms; which are being called upon to disclose more information, and to do so in a form which is more easily sliced and diced by computer. It is the eXtensible Business Reporting Language (XBRL)[1], an XML markup language for financial data, which is seen by many to be the answer to these needs. In a recently published article, XBRL is discovered to be the road to financial recovery, providing radical transparency, and letting Everyman easily ascertain the state of any company, bank or investment firm in the United States.[2]

[1] See also http://www.xbrl.org/ for more information. XBRL will be presented below in more details
[2] See http://www.wired.com/print/techbiz/it/magazine/17-03/wp_reboot (Wired article)

Similarly, in Europe XBRL is seen as a means to prevent the recurrence of financial crises by increasing transparency. There is, however, a very high barrier to transparency, namely, the large number of languages used by the European banks, investment firms and national banking supervisors in each country. Therefore also many international organisations are calling for more performing multilingual information management systems as they have to offer their information in various languages and manage information submitted to them in many different languages

A project[3] to promote the use of XBRL for disclosures made by banks and investment firms is on-going, but all of the materials it has created are only available in English. Until they are translated into the other European languages, transparency can not be granted. For this reason, it has become urgent to promote competence in all of the modern technologies of translation: terminology, translation memories and machine translation. Real transparency, however, requires more. It requires technologies to extract information from different types of documents, available in different languages, and to combine them with facts from XBRL reports, thus, further increasing the amount and quality of available multilingual information, and to transform this information into language-independent knowledge. It also requires technologies to access the knowledge in whatever language or languages the end user wants to use. The Monnet project[4] encompasses all of these technologies, and will promote competence in all of them. In a nutshell, Monnet intends to provide ontology-based multilingual access to financial data for sharing business knowledge across Europe.

In the following sections we present first the motivations and the technological apparatus of the XBRL initiative. After this we describe the general technological architecture and the main components of Monnet before introducing its XBRL-based financial and business use case.

2. XBRL Overview

XBRL, eXtensible Business Reporting Language, is an XML-based mark-up language for the exchange of business information, including financial reporting. Its use is being nowadays mandated by a growing number of regulatory bodies and stock exchanges around the world.

The widespread use of XBRL should allow regulators, analysts and investors to employ computer programs to automatically process reported information for various purposes, for example, to discover discrepancies or investment opportunities. Another important use is to integrate data from disparate accounting or Enterprise Resource Planning (ERP) systems, regardless of whether those systems are external or internal.

XBRL uses XML in a special way in order to specify the semantics of business data, its presentation, its calculation, and associated business rules, which are called *formulas*. XBRL also has its own special terminology. A set of mark-up tags for a specific purpose is called a *taxonomy*, and individual tags are called *taxonomy elements* or, alternatively, *concepts*. A computer file containing data marked up using a specific taxonomy is called an *XBRL instance*. For example, each annual report filed by a Belgian company would be an *instance*, which would be created according to the *taxonomy* for the Belgian Generally Accepted Accounting Principles (GAAP).

[3] See http://www.eurofiling.info/corepTaxonomy/taxonomy.htm

[4] MONNET (Multilingual ONtologies for NETworked knowledge))is a FP7 R&D project co-funded by the European Commission with Grant No.. 248458, See also http://www.monnet-project.eu/.

Another special feature of XBRL is that the concepts and related metadata are specified as a flat list of elements, which is separated from other information about *presentation*, *calculation*, and *business rules (formulas)*.

An example of a concept is *CurrentAssets.* Its main associated accounting metadata are:

1. It is measured at a point in time, thus its period type is *instant*, as opposed to *duration*; *duration* is the period type of a concept like *income,* which is measured over a period of time such as a year.
2. It has a balance type of *debit*, which in accounting terms means it is increased by being debited.
3. It is not an abstract concept, which means it can be used to tag items in instances.
4. It has a monetary value.

Expressed as an element in an XML schema file, an *.xsd* file, the concept might look like this (where the prefix "gaap" is used as an anonymized item):

```
<element name="CurrentAssets" id="gaap_Current_Assets" periodType="instant" balance="debit" abstract="false"
substitutionGroup="item" type="monetaryItemType"/>
```

In an instance document each value, also called a fact, is tagged with a concept as shown next.

```
<gaap:CurrentAssets contextRef="FYp0Qp0e" decimals="-6" unitRef="EUR">5255000000</gaap:CurrentAssets>
```

The *decimals* attribute value of '-6' specifies that the value is accurate to the millions. The *unitRef* specifies it is measured in Euros, and the *contextRef* points to an element that specifies the company, or *entity*, to use XBRL speak, and the time instant.

```
<context id="FYp0Qp0e">
-   <entity>
    <identifier scheme="http://www.sec.gov/CIK">0000943042</identifier>
  </entity>
- <period>
    <instant>2009-12-31</instant>
  </period>
  </context>
```

Additional information about concepts is expressed in *networks.* A network which relates concepts to each other is called a *relation network*. An example is a *presentation network,* which organizes concepts into an ordered tree, which is the basis for creating a report that is easily comprehended by analysts. There are also *resource networks,* which relate concepts to *resources*. An example of a *resource* is a label for a concept.

Separating the concepts from the labels means that an XBRL-based program can easily become multilingual. It requires only the addition of a network with labels for another language, dialect or idiolect.

Finally, XBRL taxonomies are extensible so that each reporting entity can adjust them to contain the concepts, relations and resources it needs to report on its business. The addition

of company-specific product lines, for example, would enable reporting and analysis of sales by those product lines.

The adjustment of a taxonomy is actually called *extension,* even though it may involve removing things. Extension is done without modifying the taxonomy which is being extended. It works by means of a well-defined mechanism for combining extensions with the original taxonomy. Its power is that it confers the ability to modify the conceptual model represented by the original taxonomy without modifying that taxonomy. In combination with resource networks for labels, this ability means that both the conceptual model and the associated words can be perfectly tailored to each language community or individual.

3. The Monnet project

Monnet (Multilingual ONtologies for NETworked knowledge) is a recently launched EU-funded project in the field of Language Technologies within the ICT programme[5]. Monnet is working on solutions that aim at facilitating access to on-line information across a range of languages. The initial step for Monnet is to set up infrastructures for extracting, representing and accessing knowledge across languages, using a novel combination of Semantic Web technology and automatic machine translation.

Current approaches to cross-lingual information access provide only partial solutions that address the problem in a restricted way, operating only at the document level without addressing either uniform extraction, representation, integration and querying of information across different languages and heterogeneous (textual, semi-structured, structured) data sources. Hence, the state of the art in machine translation is still far from providing multilingual services in specific domains. A key aspect of the solution Monnet is working on is reflected by the fact that the technologies deployed in the project are dealing with information at the semantic level, i.e. by abstracting away over language(s) and form of the documents, allowing for a more advanced and uniform cycle of information processing (extraction and integration) and presentation of multilingual information.

The project is validating its approach to enabling the multilingual web in the context of two use cases, one in the field of e-Government and one in the field of financial and business information. We concentrate in this paper on the second use case, for which Monnet is very happy to have as one of its member the association XBRL-Europe, which is co-defining the standards for financial and business reporting across Europe. The use case aims at enabling the search and the report creation of financial information and business service descriptions in the language of choice of the users.

Relating language-independent Knowledge Representation and language information

Monnet aims at supporting on-line information access across languages. We take as granted that the relevant information for the users is encoded in language-independent knowledge representation systems, like taxonomies and ontologies, which are supporting the uniform handling of information originally coming from different sources existing in different languages as well as the presentation of factual information in an arbitrary language.

But in order to allow ontologies to interact with multilingual text in both the analysis and the generation mode, it is necessary to model the relation that natural language expressions can have with the language independent knowledge representation systems. Most of the latter are using a "label" feature in order to encode the natural language expressions that

[5] See http://cordis.europa.eu/fp7/ict/language-technologies/home_en.html for more details

correspond to a concept. And often such labels are existing only in English, or in the language of the country for which a taxonomy or an ontology has been designed. For example in the XBRL taxonomy developed in the context of the German legislation for business reporting, we can see that the concept ID "de-gaap-ci_bs.ass.fixAss.fin.sharesInAffil.parentComp" has two associated labels: "Anteile an herrschender oder an mit Mehrheit beteiligter Gesellschaft" (in German) and "Shares in parent or in majority investor" (in English)[6].

The content of such labels are in fact just terms, which are not explicitly linked to other terms included in the labels of other concept IDs. In this a lot of information about possible linguistic realisations of concepts is left by side, and we are missing also a possible generalisation on the meaning of certain words that are used in the context of various labels within an ontology (or a taxonomy).

The semantic web, in particular with the linked data project (Bizer et al, 2009), proposes solutions that allow to re-use lexical and terminological resources by their semantic interlinking. But currently there is no standard for providing complex lexical information for ontologies promoted by the semantic web and describing the relationship between the lexicon and the ontology.

Therefore a central aspect in Monnet consists in designing and developing a model that associates linguistic information with domain semantics as defined by the corresponding (domain-specific) ontologies. Our model, which we call "Lemon" (Lexicon Model for Ontologies), is building on existing work, while extending and integrating it, in particular LMF (Francopoulo et al,2006), ISOcat (Kemps-Snijders et al, 2008), SKOS (Miles and Bechhofer, 2009), LexInfo (Buitelaar et al, 2009) and LIR (Montiel-Ponsoda et al, 2008). It is an RDF model that allows for lexical data to be shared and interlinked on the Web. Lemon is a central endeavor towards a formal model for multilingual, lexicalized knowledge representation, and it is supporting in Monnet the development of various components.

One of Monnet's roles in regard to XBRL consists in supporting the process of tailoring a taxonomy to a language community. For this we plan to either "upgrade" XBRL to an ontology so that we can combine it with the Lemon model, or at least to find a way to "transplant" Lemon onto the XBRL taxonomic way of organizing the knowledge related to business reporting.

General Architecture and Components of Monnet

The main components of Monnet, building on Lemon, are shown in the context of the general overview of the project architecture illustrated in Figure 1. On the base of the Ontology-Lexicon model, first a service for ontology lexicalization is implemented (not represented in Figure 1). This service is separating the natural language expressions used in labels of an an ontology (or a taxonomy) and enrich those automatically with linguistic information (e.g. morpho-syntax, syntax), in compliance with existing standards, like the ISO data categories proposed within ISO TC37 on "Terminology and other language resources"[7].

[6] See http://www.abra-search.com/ABRASearch.html?locale=en&taxonomy=de-gaap-ci-2010-01-31-role-labels-en-shell

[7] See http://www.isocat.org/

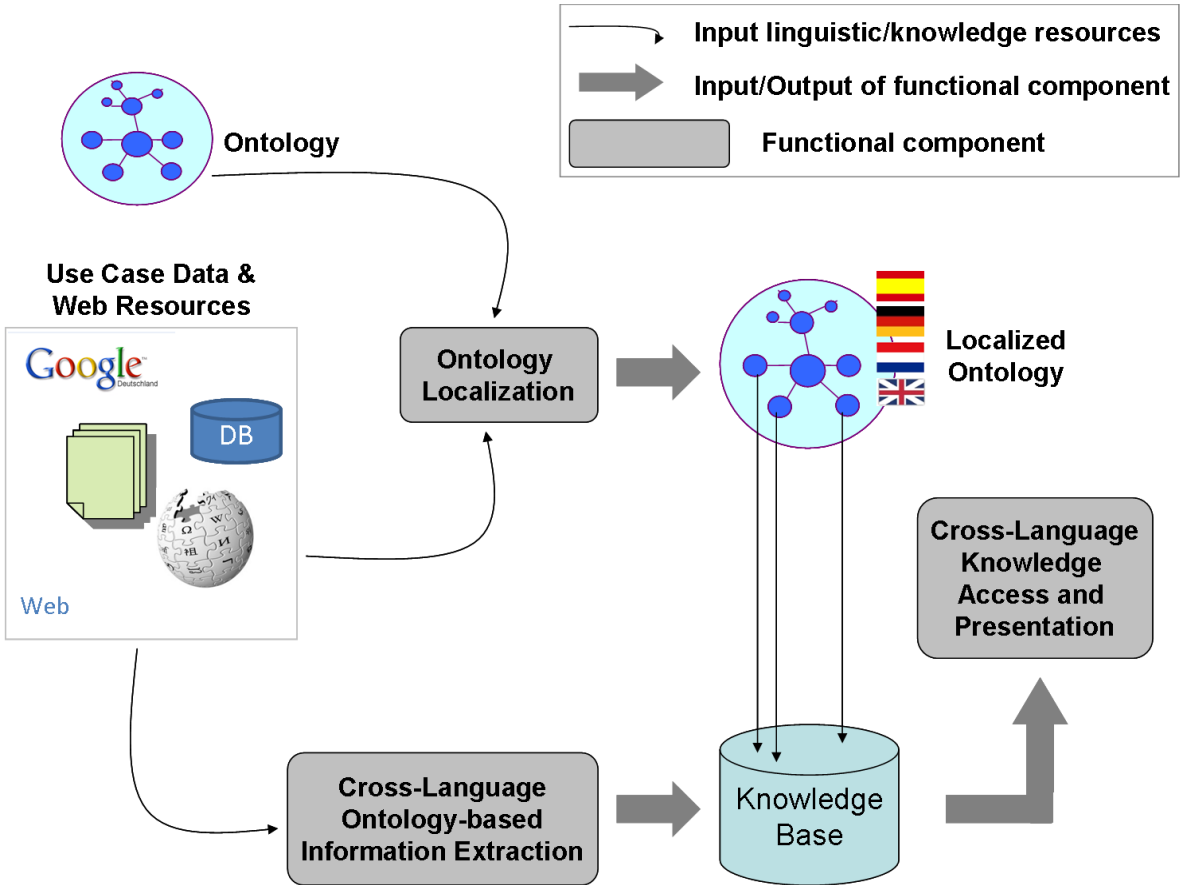

Figure 1: High-level overview of Monnet architecture and components

The Lemon model, together with the ontology lexicalization service, are the base for the three following functional components of Monnet, as can be seen in Figure 1:

- Multilingual Ontology Localisation: Creates a lexicon in a target language from a lexicon in a source language, semi-automatically
- Cross-Lingual Ontology-based Information Extraction: Uses localized ontologies for the semantic-level extraction and integration of information from text and (semi-) structured data across languages
- Cross-Lingual Knowledge Access and Presentation Framework: Uses localized ontologies for the quick customization of knowledge access systems to other languages

Multilingual Ontology Localization

Since the Multilingual Ontology Localization (MOL) is in Monnet the step following immediately the design and implementation of Lemon and the associated ontology lexicalization, and so actually attracting the main attention of the project, we give a more detailed description of this task in this paper.

MOL is the key technology on which the Monnet approach to cross-lingual information extraction, and access relies. Ontology localisation requires automatic techniques for the translation of labels used in an ontology for expressing domain concepts (in the form of classes, properties and relations). We limited the localization to the lexical layer of the ontology.

Whereas ontology localisation has received some attention lately (see Suarez-Figureoa and Gómez-Pérez, 2008), it is far from solved. The challenge is to reduce the amount of work needed to localize a given ontology by integrating automatic techniques for finding term equivalents in expressing a certain concept in different languages, e.g. a concept defining the idea of 'credit worthiness' in the financial domain can be expressed in English by the term "credit worthiness", in German by "Bonität" and in Spanish by "solvencia". To solve this task, the project recognizes the need for:

- The performance of a sound morpho-syntactic analysis that guarantees a flawless translation process and an appropriate selection among a fixed set of transfer rules that will allow the translation of compound labels from the original language to a target one
- The selection of appropriate translation techniques and methods for translating simple and complex ontology labels depending on the resources available for a certain language and domain. For example, we will rely on direct translation approaches (e.g. [Babych et al. 2007]) or word to word approaches (e.g. [Voss et al. 2008]) when authoritative and reliable multilingual dictionaries exist for the implied languages and the domain in question. When no reliable multilingual dictionaries are available, we will have to resort to other techniques such as corpus-based translation techniques (e.g [Koehn 2005]), for example, whenever parallel corpora in the implied languages and domain are available, or to statistical machine translation techniques (e.g. [López 2008] or [Chiang 2005]) combined with knowledge-based machine translation techniques (e.g. [Mohanty et al. 2007] or [Habash et al. 2006]), whenever a semantic or knowledge resource is available.
- The adaptation of the conceptualization may be necessary for cases in which the original conceptualization or ontology does not reflect the organization of a certain aspect of the domain semantics as expressed by the target language. Conceptualization adaptation requires manual inspection and goes beyond the scope of the project. We expect however that the ontology translation process will support conceptualization adaptation by appropriate identification of translation conflicts, which flags potential adaptations that need to be manually checked and updated in a separate process.

We are thus exploring approaches leading to high-quality translation, requiring only a minimal human effort to check the automatically localized ontology. Instead of producing only one possible translation, the localisation process will produce a ranked list of translation candidates which can be inspected by the human validator, achieving localisation in a semi-automatic form, which allows cost reduction and a economic interaction with the involved actors.

4. The Financial and Business Use Case

The scenario, which has been set up, is the following: A financial analyst in a certain country is looking for relevant data about companies across Europe. The relevant data might be dispersed in the following sources:

- Structured sources: balance sheets in textual and semi-structured format (Publicly available for example in German at http://www.bundesanzeiger.de/), short company profiles (e.g from European Business Register), Wikipedia Infoboxes, and XBRL instance documents (the Belgian National Bank is for example publishing on-line all the financial reports of Belgian companies that are available in XBRL)

- Semi-structured: longer company profiles, imprint information on company web pages (mainly available in Germany in line with current legal requirements)
- Unstructured: annexes to balance sheets in annual reports of companies, newspapers, specialized web pages etc. Language and legislation-specific issues in financial reporting are dealt with by the ontology-based information extraction machinery, i.e. different extraction instantiations for the different financial reporting formats/contents will be developed. The extracted information can then be harmonized and integrated on the XBRL ontology level.

The objective of this use case is to develop a prototype which allows accessing relevant data about companies originally distributed across languages and sources. The prototype will allow a financial analyst to search for data by filling in structured search forms localized to his/her own language. The results will be presented in terms of charts, diagrams, results lists etc. localized to their preferred language.

In the background, this use case will exploit the methods developed for the lexicalization service and the three main components of Monnet (see the explanatory text of Figure 1) :

- Ontology Localization: Techniques for ontology localization will be applied to the different XBRL taxonomies to localize them (mainly translating the labels) to the different languages we consider: English, German, Dutch and Spanish. The main beneficiaries here are the translators of taxonomies that can profit from the (semi-) automatic translation support provided by our ontology localization techniques. We will measure the effort reduction with respect to translating all the labels by hand here as a baseline. In localizing a given XBRL taxonomy to various languages, we will consider the taxonomies for other countries as a valuable resource and domain-specific data to guide the localization component.
- Cross-language Ontology-based Information Extraction: While some of the data we expect to exploit is already formalized according to the XBRL standards (we will refer to documents containing such data as “XBRL instance documents” as they instantiate the concepts defined in the XBRL taxonomies to convey specific factual data), there is a plethora of information that remains unstructured and dispersed across sites and languages (compare the non-exhaustive list of relevant sources above). Cross-language information extraction techniques will thus be applied to extract relevant information from documents across languages. The extracted facts will be represented in a normalized and language-independent fashion by resorting to the XBRL taxonomies.
- Cross-language Knowledge Access and Presentation: This component will allow the financial analyst to query the knowledge repository in his own language. It will use the knowledge base where the factual data is stored to answer the query and rely on the component to present this content in the language of choice of the user. While we will focus on the languages: German, English, Spanish and Dutch in the Monnet project, the approach will be able to scale to other languages as well.

The main beneficiaries of the technologies developed in this use case will thus be:

1. The translators of the XBRL taxonomies (who will profit from the effort reduction yielded by using our automatic ontology localization functionality) and also
2. The users of XBRL taxonomies (e.g. financial analysts), who will be able to see the information formalized by XBRL instance documents or extracted from unstructured resources in accordance to the XBRL taxonomies in their preferred language.

By building on semantic technologies instead of on unstructured information access approaches, we will also be able to answer aggregate queries asking for data across companies, countries etc. such as:

- Compare the revenues in 2008 of Opel (Germany), SEAT (Spain) and Ford (U.S.)
- Show financial data of European software companies with over 10.000 employees

Note that within this use case we do not aim to restructure or to align the XBRL taxonomies for different countries or legislations with each other. This is a complex task that is well beyond what can be achieved in this project. We will instead focus on localization without changing the structure of the ontologies. In case some concept is not directly translatable into the target language in a one-to-one fashion (e.g. through a single term), paraphrases will be used instead. Automatic generation of appropriate paraphrases is not in the scope of our localization component and will require the involvement of domain specialists. As we conceive ontology localization as a semi-automatic process in which a tool makes suggestions to a domain expert, involvement of experts in providing paraphrases for concepts that can not be directly translated fits well into our general approach and will be integrated into our methodology.

5. Conclusion

We presented the goals and challenges of a recently started European R&D project, which aims at supporting the collection, extraction, integration and presentation of information on the Web in a multilingual fashion, and so to support the Internet users of in having access to information in her/his own language. We gave a general overview of the technologies we started to deploy and which constitute an innovative combination of Semantic Web and machine translation and localization technologies.

As two first steps of the project, we have been I) designing and implementing a new model for the description of linguistic information of natural language expressions used in the context of labels of ontologies, and offering thus a ontology lexicalization service, which is building the base for ontology localization and ontology-based information extraction applied to a large number of different types of documents, and ii) specified 2 use cases in the field of e-Government and Financial and Business reporting, whereas we focused in this paper in describing the business use case.

Acknowledgment

The work presented in this paper is currently under progress within the R&D project “Monnet”, which is co-funded by the European Union under Grant No. 248458.

References

Babych B. , A. Hartley, and S. Sharoff (2007). Translating from under-resourced languages: comparing direct transfer against pivot translation. MT Summit XI, 10-14. Copenhagen.

Buitelaar P, Cimiano P, Haase P, Sintek M (2009). Towards linguistically grounded ontologies. The Semantic Web: Research and Applications pp 111-125

D. Chiang (2005). A Hierarchical Phrase-Based Model for Statistical Machine Translation. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)

Chiarcos C (2010). Grounding an Ontology of Linguistic Annotations in the Data Category Registry. LREC10 W4 p 37.

Francopoulo G, George M, Calzolari N, Monachini M, Bel N, Pet M, Soria C (2006). Lexical markup framework (LMF). Proceedings of LREC2006 pp 233-236

Habash N., C. Mah, S. Imran, R. Calistri-Yeh, and P.Sheridan (2006). Design, construction and validation of an Arabic-English conceptual interlingua for cross-lingual information retrieval. LREC-2006: Fifth International Conference on Language Resources and Evaluation.

Isaac A, Phipps J, Rubin D (2009). SKOS Use Cases and Requirements. URL http: //www.w3.org/TR/2009/NOTE-skos-ucr-20090818/.

Kemps-Snijders M, Windhouwer M, Wittenburg P, Wright S (2008). ISOcat: Corralling data categories in the wild. In: Proceedings of the International Conference on Language Resources and Evaluation, Marrakech, Morocco.

P. Koehn, F.J. Och, and D. Marcu (2003). Statistical phrase based translation. In Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of *the North American Chapter of the Association of Computational Linguistic.*.

Mohanty R.K., M. Krishna Prasad, L. Narayanaswamy, P. Bhattacharyya (2007). Semantically relatable sequences in the context of interlingua based machine translation. ICON-2007. 5th International Conference on Natural Language Processing, IIIT Hyderabad, India.

Montiel-Ponsoda E, de Cea G, Gomez-Perez A, Peters W (2008). Modelling multi-linguality in ontologies. In: Proceedings of the 22nd International Conference on Computational Linguistics, Coling.

Romary L (2010). Standardization of the formal representation of lexical information for NLP. Dictionaries An International Encyclopedia of Lexicography Supplementary volume: Recent developments with special focus on computational lexicography.

Shadbolt N, Hall W, Berners-Lee T (2006) The semantic web revisited. IEEE intelligent systems 21(3):96-101.

Suárez-Figueroa M.C., A. Gómez-Pérez (2008). First Attempt towards a Standard Glossary of Ontology Engineering Terminology. Proceedings of the 8thInternational Conference on Terminology and Knowledge Engineering (TKE2008), Copenhagen, August 2008.

Voss C.R. , M. Aguirre, J. Micher, R. Chang, J. Laoudi, and R. Hobbs (2008). Boosting performance of weak MT engines automatically: using MT output to align segments & build statistical post-editors. EAMT 2008: 12th annual conference of the European Association for Machine Translation, September 22 & 23, 2008, Hamburg, Germany.