3rd Workshop on the Multilingual Semantic Web

In conjunction with the International Semantic Web Conference (ISWC2012)

Sponsored by the EU funded project Monnet: Multilingual Ontologies for Networked Knowledge (http://www.monnet-project.eu)

Boston, USA, November 11th, 2012

About

The vision of the Multilingual Semantic Web workshop series is the creation of a Semantic Web where semantically structured information can be aligned, integrated and used across languages. The workshops are concerned with research questions on how current Semantic Web infrastructure can and should be extended to support this vision.

Ontologies and linked data vocabularies are defined often in one language only (English), with a biased semantics and a corresponding world view. An infrastructure should be in place for defining ontologies and vocabularies in multiple languages with a transparent semantics across them. Current Semantic Web representation languages (RDF, OWL, SKOS) are limited in regard of the representation of natural language semantics, leaving much of the semantics hidden in textual web content out of scope for the developing Web of Data.

NLP and machine learning for Linked Data can benefit from exploiting linguistic resources such as annotated corpora, wordnets etc. if they are themselves formally represented and linked by use of Linked Data principles. In addressing such research questions, the workshop aims at providing a forum for researchers at the intersection of NLP, multilingual information access, Linked Data and the Semantic Web to exchange ideas on realizing the Multilingual Semantic Web.

Motivation

Although knowledge processing on the Semantic Web is inherently language-independent, human interaction with semantically structured and linked data will remain inherently language-based as it often requires text or speech input – in many different languages. Semantic Web development will therefore be increasingly concerned with knowledge extraction, integration and interaction in multiple languages, making multilinguality an emerging challenge to the global advance of Semantic Web and linked data use and development across language communities around the world.

The 3rd workshop on Multilingual Semantic Web has more focus on the underlying multilingual web infrastructure as well as the linguistic annotation needed for multilingual knowledge extraction, integration and interaction.

The workshop will be supported by the following W3C and ISO working groups:

- Ontology-Lexica W3C Community Group
- MultilingualWeb-LT WG
- ISO-Space project
- ISO-TimeML

We kindly acknowledge the European Union for its support through the research grant for Monnet (FP7-248458) and the Science Foundation Ireland through Lion2 (SFI/08/CE/I1380)

Program Committee

Lupe Aguado de Cea, Informatics & Applied Linguistics depts., Univ. Politécnica de Madrid, Spain

Roberto Basili, Artificial Intelligence group, Department of Computer Science, Tor Vergata University of Rome, Italy

Gerhard Budin, Center of Translation Studies, University of Vienna, Austria

Nicoletta Calzolari, Inst. of Computational Linguistics, NRC, Pisa, Italy

Christian Chiarcos, Germany

Thierry Declerck, Language Technology Lab, DFKI GmbH, Germany

Gerard De Melo, AI/FrameNet group, International Computer Science Institute (ICSI), UC Berkeley, USA

Bo Fu, Computer Human Interaction & Software Engineering Lab, Computer Science dept., Univ. of Victoria, BC, Canada

Aldo Gangemi, Semantic Technology Lab, CNR Institute of Cognitive Sciences and Technology, Rome, Italy

Jorge Gracia, Ontology Engineering Group, AI dept., Universidad Politécnica de Madrid (UPM), Spain

Judith Eckle-Kohler, Ubiquitous Knowledge Processing Lab, Computer Science dept., Tech. Univ. Darmstadt, Germany

Yoshihiko Hayashi, Graduate School of Language and Culture, Osaka University, Japan

Sebastian Hellman, Business Information Systems, Univ. of Leipzig, Germany

Graeme Hirst, Computer Science dept., University of Toronto, Canada

Antoine Isaac, Department of Computer Science, Vrije Universiteit, The Netherlands

Nancy Ide, Department of Computer Science, Vassar College, USA

Hitoshi Isahara, Toyohashi Institute of Technology, Japan

Zornitsa Kozareva, Information Sciences Institute, University of Southern California, USA

John McCrae, Semantic Computing Group, CITEC, University of Bielefeld, Germany

Elena Montiel-Ponsoda, Ontology Engineering Group, AI dept., Univ. Politécnica de Madrid, Spain

Roberto Navigli, Dept. of Computer Science, Linguistic Computing Laboratory, Sapienza University of Rome, Italy

Sergei Nirenburg, Computer Science & Electrical Engineering, University of Maryland, USA

Jong-Hoon Oh, Information Analysis Laboratory, Universal Communication Research Institute, NICT, Japan

Thierry Poibeau, LaTTiCe (Langues, Textes, Traitements informatiques et Cognition), CNRS, France

Laurette Pretorius, Computer Science dept., University of South Africa, South-Africa

Martin Volk, Institut für Computerlinguistik, Universität Zürich, Switzerland

Piek Vossen, Computational Lexicology, Vrije Universiteit, The Netherlands

Organizing Committee

Dr. Paul Buitelaar
Unit for Natural Language Processing, Digital Enterprise Research Institute (DERI)
National University of Ireland, Galway
paul.buitelaar@deri.org
http://www.paulbuitelaar.net/

Prof. Philipp Cimiano
Semantic Computing Group, Center of Excellence Cognitive Interaction Technology
University of Bielefeld, Germany
cimiano@cit-ec.uni-bielefeld.de
http://www.cimiano.de

Dr. David Lewis
Knowledge and Data Engineering Group, School of Computer Science and Statistics
Trinity College Dublin, Ireland
dave.lewis@cs.tcd.ie
http://www.scss.tcd.ie/dave.lewis/

Prof. James Pustejovsky
Department of Computer Science, Volen Center for Complex Systems
Brandeis University, Waltham, MA, USA
jamesp@cs.brandeis.edu
http://pages.cs.brandeis.edu/~jamesp/

Prof. Felix Sasaki
W3C Fellow & Language Technology Lab, DFKI GmbH
Berlin, Germany
Felix.Sasaki@dfki.de
http://www.w3.org/People/fsasaki/

Table of Contents

BabelNet goes to the (Multilingual) Semantic Web

Roberto Navigli

Sapienza University of Rome, Via Salaria, 113 – 00198 Roma Italy,
navigli@di.uniroma1.it

Abstract. BabelNet is a very large, wide-coverage multilingual ontology. This resource is created by linking the largest multilingual Web encyclopedia – i.e., Wikipedia – to the most popular computational lexicon – i.e., WordNet. The integration is performed via an automatic mapping and by filling in lexical gaps in resource-poor languages with the aid of Machine Translation. The result is an "encyclopedic dictionary" that provides babel synsets, i.e., concepts and named entities lexicalized in many languages and connected with large amounts of semantic relations. BabelNet is available online at http://www.babelnet.org.
In this paper we present a first attempt at encoding BabelNet for the multilingual Semantic Web.

Keywords: lexicalized ontologies, semantic networks, multilinguality, lexical semantics

1 Introduction

In the information society, lexical knowledge is a key skill for understanding and decoding an ever-changing world. Indeed, lexical knowledge is an essential component not only for human understanding of text, but it is also indispensable for the creation of the multilingual Semantic Web. Unfortunately, however, building such lexical knowledge resources manually is an onerous task requiring dozens of years – and what is more it has to be repeated from scratch for each new language. On top of this, it is becoming increasingly critical that existing resources be published as Linked Open Data (LOD), so as to foster integration, interoperability and reuse on the Semantic Web [3].

Thus, lexical resources provided in RDF format [4] can contribute to the creation of the so-called Linguistic Linked Open Data (LLOD, see Figure 1), a vision fostered by the Open Linguistic Working Group (OWLG)[1] in which part of the Linked Open Data cloud is made up of interlinked linguistic resources [2]. The multilinguality aspect is key to this vision, in that it enables Natural Language Processing tasks which are not only cross-lingual, but also independent of the language of the user input and the linked data exploited to perform the task.

[1] http://linguistics.okfn.org

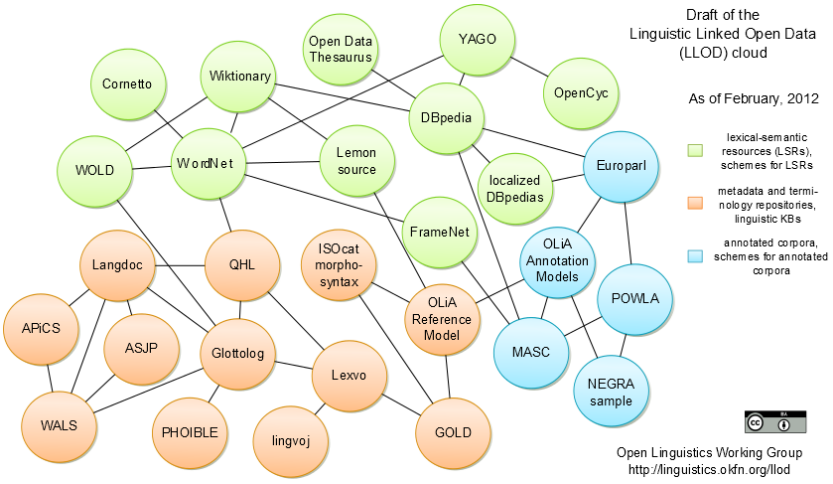

Fig. 1. Open Linguistics Working Group (2012), The Linguistic Linked Open Data cloud diagram (draft), version of February 2012, `http://linguistics.okfn.org/llod`.

This paper provides a contribution to the LLOD vision by presenting a first encoding of BabelNet in RDF. BabelNet (`http://www.babelnet.org`) is a very large multilingual semantic network obtained as a result of a novel integration and enrichment methodology. This resource is created by linking the largest multilingual Web encyclopedia – i.e., Wikipedia – to the most popular computational lexicon – i.e., WordNet [6]. The integration is performed via an automatic mapping and by filling in lexical gaps in resource-poor languages with the aid of Machine Translation (MT). The result is an "encyclopedic dictionary" that provides babel synsets, i.e., concepts and named entities lexicalized in many languages and connected with large amounts of semantic relations.

While the LOD is centered around DBPedia [1], the largest "hub" of Linked Data which provides wide coverage of Named Entities, BabelNet focuses both on word senses and on Named Entities in many languages. Therefore, its aim is to provide full lexicographic and encyclopedic coverage. Compared to YAGO [11], BabelNet integrates WordNet and Wikipedia by means of a mapping strategy based on a disambiguation algorithm, and provides additional lexicalizations resulting from the application of MT.

In the next Section we introduce BabelNet and briefly illustrate its features. Then, in Section 3 we provide statistics and in Section 4 we describe the RDF encoding of BabelNet. Finally, we give some conclusions in Section 5.

2 BabelNet

BabelNet [8] encodes knowledge as a labeled directed graph $G = (V, E)$ where V is the set of nodes – i.e., concepts such as balloon and named entities such as Montgolfier brothers – and $E \subseteq V \times R \times V$ is the set of edges connecting

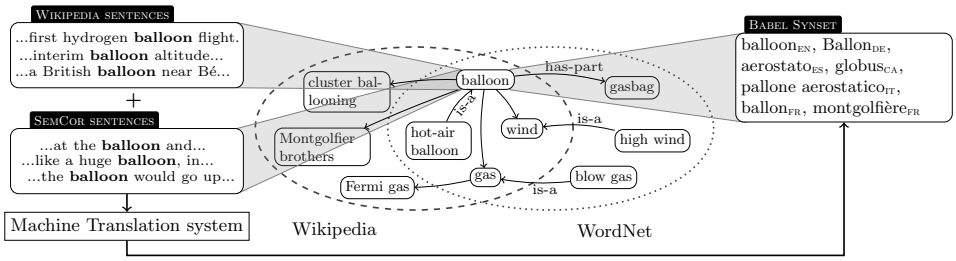

Fig. 2. An overview of BabelNet (nodes are labeled with English lexicalizations only): unlabeled edges are extracted from Wikipages (e.g., Balloon (aircraft) links to Montgolfier brothers), labeled edges come from WordNet (e.g., balloon^1_n *has-part* gasbag^1_n).

pairs of concepts (e.g., balloon *is-a* lighter-than-air craft). Each edge is labeled with a semantic relation from R, e.g., $\{$*is-a*, *part-of*, ..., $\epsilon\}$, where ϵ denotes an unspecified semantic relation. Each node $v \in V$ contains a set of lexicalizations of the concept for different languages, e.g., $\{$ balloon$_{\text{EN}}$, Ballon$_{\text{DE}}$, pallone aerostatico$_{\text{IT}}$, ..., montgolfière$_{\text{FR}}$ $\}$. We call such multilingually lexicalized concepts *Babel synsets*. Concepts and relations in BabelNet are harvested from the largest available semantic lexicon of English, WordNet, and a wide-coverage collaboratively-edited encyclopedia, Wikipedia. In order to build the BabelNet graph, we collect at different stages:

a. from WordNet, all available word senses (as concepts) and all the lexical and semantic pointers between synsets (as relations);
b. from Wikipedia, all encyclopedic entries (i.e., Wikipages, as concepts) and semantically unspecified relations from hyperlinked text.

An overview of BabelNet is given in Figure 2. The excerpt highlights that WordNet and Wikipedia can overlap both in terms of concepts and relations: accordingly, in order to provide a *unified resource*, we merge the intersection of these two knowledge sources. Next, to enable multilinguality, we collect the lexical realizations of the available concepts in different languages. Finally, we connect the multilingual Babel synsets by establishing semantic relations between them. Thus, our methodology consists of three main steps:

1. We **combine WordNet and Wikipedia** by automatically acquiring a mapping between WordNet senses and Wikipages. This avoids duplicate concepts and allows their inventories of concepts to complement each other.
2. We **harvest multilingual lexicalizations** of the available concepts (i.e., Babel synsets) by using (a) the human-generated translations provided by Wikipedia (the so-called *inter-language* links), as well as (b) a machine translation system to translate occurrences of the concepts within sense-tagged corpora.

Language	Lemmas	Synsets	Word senses
English	5,938,324	3,032,406	6,550,579
Catalan	3,518,079	2,214,781	3,777,700
French	3,754,079	2,285,458	4,091,456
German	3,602,447	2,270,159	3,910,485
Italian	3,498,948	2,268,188	3,773,384
Spanish	3,623,734	2,252,632	3,941,039
Total	23,935,611	3,032,406	26,044,643

Table 1. Number of lemmas, synsets and word senses in the 6 languages currently covered by BabelNet.

3. We **establish relations between Babel synsets** by collecting all relations found in WordNet, as well as all wikipedias in the languages of interest: in order to encode the strength of association between synsets, we compute their degree of correlation using a measure of relatedness based on the Dice coefficient.

3 Statistics

In this section we provide statistics for BabelNet 1.0.1, obtained by applying the construction methodology briefly described in the previous Section and detailed in [8].

3.1 WordNet-Wikipedia mapping

The overall mapping contains 89,226 pairs of Wikipages and WordNet senses they map to, covers 52% of the noun senses in WordNet, with an accuracy of about 82% estimated on a random sample of 1,000 items.

3.2 Lexicon

BabelNet currently covers 6 languages, namely: English, Catalan, French, German, Italian and Spanish. Its lexicon includes lemmas which denote both concepts (e.g., balloon) and named entities (e.g., Montgolfier brothers). The second column of Table 1 shows the number of lemmas for each language. The lexicons have the same order of magnitude for the 5 non-English languages, whereas English shows larger numbers due to the lack of inter-language links and annotated sentences for many terms, which prevents our construction approach from providing translations.

In Table 2 we report the number of monosemous and polysemous words divided by part of speech. Given that we work with nominal synsets only, the numbers for verbs, adjectives and adverbs are the same as in WordNet 3.0. As for nouns, we observe a very large number of monosemous words (almost 23 million), but also a large number of polysemous words (more than 1 million).

POS	Monosemous words	Polysemous words
Noun	22,763,265	1,134,857
Verb	6,277	5,252
Adjective	16,503	4,976
Adverb	3,748	733
Total	22,789,793	1,145,818

Table 2. Number of monosemous and polysemous words by part of speech (verbs, adjectives and adverbs are the same as in WordNet 3.0).

		English	Catalan	French	German	Italian	Spanish	Total
English WordNet		206,978	-	-	-	-	-	206,978
Wikipedia	pages	2,955,552	123,101	524,897	506,892	404,153	349,375	4,863,970
	redirections	3,388,049	105,147	617,379	456,977	217,963	404,009	5,189,524
	translations	-	3,445,273	2,844,645	2,841,914	3,046,323	3,083,365	15,261,520
WordNet	monosemous	-	97,327	97,680	97,852	98,089	97,435	488,383
	SemCor	-	6,852	6,855	6,850	6,856	6,855	34,268
Total		6,550,579	3,777,700	4,091,456	3,910,485	3,773,384	3,941,039	26,044,643

Table 3. Composition of Babel synsets: number of synonyms from the English WordNet, Wikipedia pages and translations, as well as translations of WordNet's monosemous words and SemCor's sense annotations.

Both numbers are considerably larger than in WordNet, because – as remarked above – words here denote both concepts (mainly from WordNet) and named entities (mainly from Wikipedia).

3.3 Concepts

BabelNet contains more than 3 million concepts, i.e., Babel synsets, and more than 26 million word senses (regardless of their language). In Table 1 we report the number of synsets covered for each language (third column) and the number of word senses lexicalized in each language (fourth column). 72.3% of the Babel synsets contain lexicalizations in all 6 languages and the overall number of word senses in English is much higher than those in the other languages (owing to the high number of synonyms available in the English WordNet synsets). Each Babel synset contains 8.6 synonyms, i.e., word senses, on average, in any language. The number of synonyms per synset for each language individually ranges from a maximum 2.2 for English to a minimum 1.7 for Italian, with an average of 1.8 synonyms per language.

In Table 3 we show for each language the number of word senses obtained directly from WordNet, Wikipedia pages and redirections, as well as Wikipedia and WordNet translations.

3.4 Relations

We now turn to relations in BabelNet. Relations come either from Wikipedia hyperlinks (in any of the covered languages) or WordNet. All our relations are

	English	Catalan	French	German	Italian	Spanish	Total
WordNet	364,552	-	-	-	-	-	364,552
WordNet glosses	617,785	-	-	-	-	-	617,785
Wikipedia	50,104,884	978,006	5,613,873	5,940,612	3,602,395	3,411,612	69,651,382
Total	51,087,221	978,006	5,613,873	5,940,612	3,602,395	3,411,612	70,633,719

Table 4. Number of lexico-semantic relations harvested from WordNet, WordNet glosses and the 6 wikipedias.

English	WordNet	Large tough nonrigid bag filled with gas or heated air.
	Wikipedia	A balloon is a type of aircraft that remains aloft due to its buoyancy.
German		Ein Ballon ist eine nicht selbsttragende, gasdichte Hülle, die mit Gas gefüllt ist und über keinen Eigenantrieb verfügt.
Italian		Un pallone aerostatico è un tipo di aeromobile, un aerostato che si solleva da terra grazie al principio di Archimede.
Spanish		Un aerostato, o globo aerostático, es una aeronave no propulsada que se sirve del principio de los fluidos de Arquímedes para volar, entendiendo el aire como un fluido.

Table 5. Glosses for the Babel synset referring to the concept of balloon as aircraft'.

semantic, in that they connect Babel synsets (rather than senses), however the relations obtained from Wikipedia are unlabeled.[2] In Table 4 we show the number of lexico-semantic relations from WordNet, WordNet glosses and the 6 wikipedias used in our work. We can see that the major contribution comes from the English Wikipedia (50 million relations) and Wikipedias in other languages (a few million relations, depending on their size in terms of number of articles and links therein).

3.5 Glosses

Each Babel synset naturally comes with one or more glosses (possibly available in many languages). In fact, WordNet provides a textual definition for each English synset, while in Wikipedia a textual definition can be reliably obtained from the first sentence of each Wikipage[3]. Overall, BabelNet includes 4,683,031 glosses (2,985,243 of which are in English). In Table 5 we show the glosses for the Babel synset which refers to the concept of balloon as 'aircraft'.

3.6 Sense-tagged corpus

BabelNet also includes a sense-tagged corpus containing the sentences input to the Machine Translation system. The corpus, called BabelCor, is built by collect-

[2] In a future release of the resource we plan to perform an automatic labeling based on work in the literature. See [7] for recent work on the topic.

[3] "The article should begin with a short declarative sentence, answering two questions for the nonspecialist reader: *What (or who) is the subject?* and *Why is this subject notable?*", extracted from `http://en.wikipedia.org/wiki/Wikipedia:Writing_better_articles`. This simple, albeit powerful, heuristic has been previously used successfully to construct a corpus of definitional sentences [10] and learn a definition and hypernym extraction model [9].

ing from SemCor and Wikipedia those sentences which contain an occurrence of a polysemous word labeled with a WordNet sense (in SemCor) or hyperlinked to a Wikipage (in Wikipedia). A frequency threshold of at least 3 sentences per sense is used in order to make sure that meaningful statistics are computed from the MT system's output, thus ensuring precision. As a result, BabelCor contains almost 2 million sentences (1,986,557 in total, of which 46,155 from SemCor and 1,940,402 from Wikipedia), which provide sense-annotated data for 330,993 senses contained in BabelNet (6,856 from WordNet and 324,137 from Wikipedia).

4 BabelNet in RDF

We now introduce a first RDF encoding of BabelNet. Other encodings, including one in the Lemon RDF model [5], will be made available online soon.

4.1 Babel synsets in RDF

An excerpt of the RDF Babel synset representation follows:

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:bn10schema="http://lcl.uniroma1.it/babelnet/bn10/schema/"
    xmlns:bn10instances="http://lcl.uniroma1.it/babelnet/bn10/instance/">
  ...
  <bn10schema:BabelSynset
    rdf:about="http://lcl.uniroma1.it/babelnet/bn10/instance/bn:00008187n">
    <bn10schema:pos>NOUN</bn10schema:pos>
    <bn10schema:source>WIKIWN</bn10schema:source>
    <bn10schema:babelSynsetId>bn:00008187n</bn10schema:babelSynsetId>
    <bn10schema:mainSense>balloon#n#1</bn10schema:mainSense>
    <bn10schema:semanticallyRelated>
      <bn10schema:BabelSynset
      rdf:about="http://lcl.uniroma1.it/babelnet/bn10/instance/bn:02955250n">
        <bn10schema:mainSense>WIKI:EN:Montgolfier_brothers</bn10schema:mainSense>
      </bn10schema:BabelSynset>
    </bn10schema:semanticallyRelated>
    <bn10schema:hypernym>
      <bn10schema:BabelSynset
      rdf:about="http://lcl.uniroma1.it/babelnet/bn10/instance/bn:00051149n">
        <bn10schema:mainSense>lighter-than-air_craft#n#1</bn10schema:mainSense>
      </bn10schema:BabelSynset>
    </bn10schema:hypernym>
  </bn10schema:BabelSynset>
  <bn10schema:BabelSynset
    rdf:about="http://lcl.uniroma1.it/babelnet/bn10/instance/bn:01631774n">
    <bn10schema:pos>NOUN</bn10schema:pos>
    <bn10schema:source>WIKI</bn10schema:source>
    <bn10schema:babelSynsetId>bn:01631774n</bn10schema:babelSynsetId>
```

```
    <bn10schema:mainSense>WIKI:EN:First_flying_machine</bn10schema:mainSense>
  </bn10schema:BabelSynset>
  <bn10schema:BabelSynset
    rdf:about="http://lcl.uniroma1.it/babelnet/bn10/instance/bn:02955250n">
    <bn10schema:pos>NOUN</bn10schema:pos>
    <bn10schema:source>WIKI</bn10schema:source>
    <bn10schema:babelSynsetId>bn:02955250n</bn10schema:babelSynsetId>
    <bn10schema:mainSense>WIKI:EN:Montgolfier_brothers</bn10schema:mainSense>
  </bn10schema:BabelSynset>
  ...
</rdf:RDF>
```

The excerpt above encodes the three Babel synsets for the concepts of balloon (in the sense of aircraft), first flying machine and Montgolfier brothers. The `<pos>` tag provides the part of speech tag of the synset, the `<source>` tag describes the source from which the synset was obtained (WN for WordNet, WIKI for Wikipedia, WIKIWN for the intersection between the two resources), `<babelSynsetId>` provides the numeric id of the synset, and `<mainSense>` provides the main sense (either from WordNet or Wikipedia) which univocally identifies the Babel synset.

The first Babel synset listed above, i.e., the concept of balloon (`bn:00008187n`), is semantically related to the Montgolfier brothers (`bn:02955250n`), among others, as encoded by the `semanticallyRelated` relation, and is a lighter-than-air craft (`bn:00051149n`), as encoded by the `hypernym` relation.

4.2 Babel senses in RDF

An excerpt of the RDF Babel sense representation follows:

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:bn10schema="http://lcl.uniroma1.it/babelnet/bn10/schema/"
    xmlns:bn10instances="http://lcl.uniroma1.it/babelnet/bn10/instance/">
  ...
  <bn10schema:BabelSense
    rdf:about="http://lcl.uniroma1.it/babelnet/bn10/instance/
                    Balloon_(aircraft)-EN@bn:00008187n">
    <bn10schema:babelSynsetId>bn:00008187n</bn10schema:babelSynsetId>
    <bn10schema:lang>EN</bn10schema:lang>
    <bn10schema:source>WIKI</bn10schema:source>
    <bn10schema:pos>NOUN</bn10schema:pos>
    <bn10schema:lemma>Balloon</bn10schema:lemma>
  </bn10schema:BabelSense>
  <bn10schema:BabelSense
    rdf:about="http://lcl.uniroma1.it/babelnet/bn10/instance/
                    Ballongas-DE@bn:00008187n">
    <bn10schema:babelSynsetId>bn:00008187n</bn10schema:babelSynsetId>
    <bn10schema:lang>DE</bn10schema:lang>
    <bn10schema:source>WIKIRED</bn10schema:source>
```

```
    <bn10schema:pos>NOUN</bn10schema:pos>
    <bn10schema:lemma>Ballongas</bn10schema:lemma>
  </bn10schema:BabelSense>
  <bn10schema:BabelSense
    rdf:about="http://lcl.uniroma1.it/babelnet/bn10/instance/
                    Ballon-DE@bn:00008187n">
    <bn10schema:babelSynsetId>bn:00008187n</bn10schema:babelSynsetId>
    <bn10schema:lang>DE</bn10schema:lang>
    <bn10schema:source>WIKI</bn10schema:source>
    <bn10schema:pos>NOUN</bn10schema:pos>
    <bn10schema:lemma>Ballon</bn10schema:lemma>
  </bn10schema:BabelSense>
  <bn10schema:BabelSense
    rdf:about="http://lcl.uniroma1.it/babelnet/bn10/instance/
                    ballon-FR@bn:00008187n">
    <bn10schema:babelSynsetId>bn:00008187n</bn10schema:babelSynsetId>
    <bn10schema:source>WNTR</bn10schema:source>
    <bn10schema:lang>FR</bn10schema:lang>
    <bn10schema:source>WIKITR</bn10schema:source>
    <bn10schema:pos>NOUN</bn10schema:pos>
    <bn10schema:lemma>ballon</bn10schema:lemma>
  </bn10schema:BabelSense>
  <bn10schema:BabelSense
    rdf:about="http://lcl.uniroma1.it/babelnet/bn10/instance/
              Pallone_aerostatico-IT@bn:00008187n">
    <bn10schema:babelSynsetId>bn:00008187n</bn10schema:babelSynsetId>
    <bn10schema:lang>IT</bn10schema:lang>
    <bn10schema:source>WIKI</bn10schema:source>
    <bn10schema:pos>NOUN</bn10schema:pos>
    <bn10schema:lemma>pallone aerostatico</bn10schema:lemma>
  </bn10schema:BabelSense>
  ...
</rdf:RDF>
```

where `<lang>` represents the language in which the sense is lexicalized, `<source>` is the source from which the sense is obtained (WN for WordNet, WNTR or WIKITR for translations of WordNet- or Wikipedia-annotated text, WIKIRED for a Wikipedia redirection, etc.), `<pos>` is the part of speech tag of the sense, and `<lemma>` specifies the lexicalization for the sense.

5 Conclusions

The Web of Data is in need for multilingual lexicalizations for Linked Open Data. This vision of a Linguistic Linked Open Data (LLOD) has recently been promoted, among others, by the Open Linguistic Working Group as well as other researchers [3]. BabelNet [8] – an ongoing project[4] at the Sapienza Linguistic

[4] Developed in the context of the MultiJEDI ERC Starting Grant: `http://lcl.uniroma1.it/multijedi`.

Computing Laboratory[5] – fits this vision by providing multilingual lexicalizations in RDF for millions of concepts, called Babel synsets, as well as a huge network of semantic relations between them. BabelNet currently covers 6 languages, but is continuously expanded with new information and languages.

Future steps include, among others, the integration of a mapping between BabelNet and other linguistic resources which are already part of the LLOD, such as DBPedia.

Acknowledgments

The author gratefully acknowledges the support of the ERC Starting Grant MultiJEDI No. 259234. The author wishes to thank Giovanni Stilo for his help with the RDF encoding of BabelNet.

References

1. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: Dbpedia - a crystallization point for the web of data. Journal of Web Semantics 7(3), 154–165 (2009)
2. Chiarcos, C., Hellmann, S., Nordhoff, S.: Towards a linguistic linked open data cloud: The Open Linguistics Working Group. TAL 52(3), 245–275 (2011)
3. Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., McCrae, J.: Challenges for the multilingual web of data. J. Web Sem. 11, 63–71 (2012)
4. Lassila, O., Swick, R.R.: Resource description framework (rdf) model and syntax specication. In: Technical report, World Wide Web Consortium (1999)
5. McCrae, J., Spohr, D., Cimiano, P.: Linking lexical resources and ontologies on the Semantic Web with Lemon. In: Proceedings of the 8th Extended Semantic Web Conference (ESWC). pp. 245–259. Heraklion, Crete, Greece (2011)
6. Miller, G.A., Beckwith, R., Fellbaum, C.D., Gross, D., Miller, K.: WordNet: an online lexical database. International Journal of Lexicography 3(4), 235–244 (1990)
7. Moro, A., Navigli, R.: WiSeNet: Building a Wikipedia-based semantic network with ontologized relations. In: Proceedings of the 21^{st} ACM Conference on Information and Knowledge Management (CIKM 2012). Maui, HI, USA (2012)
8. Navigli, R., Ponzetto, S.P.: BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence 193, 217–250 (2012)
9. Navigli, R., Velardi, P.: Learning Word-Class Lattices for definition and hypernym extraction. In: Proceedings of the 48^{th} Annual Meeting of the Association for Computational Linguistics (ACL). pp. 1318–1327. Uppsala, Sweden (2010)
10. Navigli, R., Velardi, P., Ruiz-Martínez, J.M.: An annotated dataset for extracting definitions and hypernyms from the web. In: Proceedings of the 7th International Conference on Language Resources and Evaluation, Valletta, Malta, 19–21 May 2010 (2010)
11. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A large ontology from Wikipedia and WordNet. Journal of Web Semantics 6(3), 203–217 (2008)

[5] `http://lcl.uniroma1.it`

Towards the Generation of Semantically Enriched Multilingual Components of Ontology Labels

Thierry Declerck and Dagmar Gromann

DFKI GmbH, Language Technology Department,
Stuhlsatzenhausweg 3, D-66123 Saarbruecken, Germany
declerck@dfki.de
Vienna University of Economics and Business
Nordbergstrasse 15, 1090 Vienna, Austria
dgromann@wu.ac.at

Abstract. Ontologies often contain multilingual textual information in annotation properties, such as `rdfs:label` and `rdfs:comment`. While the motivation for using such annotation properties is to provide a human readable description of abstract conceptualization of the domain, we notice that the importance of appropriate natural language use and representation is often neglected. The same can be observed with resources on the Web, such as multilingual taxonomies. Terms often lack consistency and completeness, hampering also an accurate automated natural language processing of such text. We propose a pattern-based transformation of terms in labels, thereby also supporting a multilingual alignment of (sub)components of labels. The source data for our approach is an ontology we derived from an industry classification taxonomy, which we improve as regards consistency and completeness and apply to the process of lexicalization.

Keywords: Ontology Labels, Multilingualism, Terms and Sub-Terms

1 Introduction

Nowadays, it has been increasingly realized that the process of ontology construction is inevitably linked to natural language and related to this development multilingualism is progressively gaining center stage in ontology engineering. There are various possibilities to add natural language strings to ontologies. These strings can be part of RDF URI references, identifying ontological resources (e.g. natural language string used in `rdf:ID`), a fragment (e.g. natural language string in `rdf:about` statements) or marking empty property elements (kind of leaf nodes in a graph, using the `rdf:resource` statement). Natural language strings equally represent the content of the RDF annotation properties `rdfs:label` and `rdfs:comment`, which provide information on ontological resources in a human-readable format.

Herein, we focus on the content of annotation properties. This choice has been partially motivated by the fact that these properties qualify for the inclusion of

terminological information, which can be realized in form of longer natural language strings. Additionally, labels and comments locally support multilinguality by means of language tags of RDF literals, i.e., `xml:lang`, whereas this is not the case for RDF URI references.

Analyzing the content of annotation properties in multilingual ontologies, we registered that their realization frequently hampers an accurate automatic linguistic and semantic processing. This type of processing is vital to a large number of ontology-based tasks, such as machine translation, information extraction, cross-lingual ontology mapping. Thus, we investigate if and how cross-lingual preprocessing and linguistic harmonization of terms in ontology labels can be of avail for such processing. At the same time, these initial steps support a multilingual alignment of subcomponents of labels, leading to more fine-grained multilingual resources associated with ontology elements.

Our experimental results are based on the analysis of labels and comments of an ontology we derived from the Global Industry Classification Standard (GICS) taxonomy[1] in English, German, and Spanish. The GICS taxonomy consists of four meta-levels, namely, sector, industry group, industry, sub-industry. These four categories represent the top nodes of the ontology. Each leaf node, i.e., each sub-industry, contains a detailed definition. All classes are indexed by integers, which also indicate the hierarchical structure of the taxonomy: the descending line "10" (Energy), "1010" (Energy), "101010" (Energy Equipment & Services) and "10101010" (Oil & Gas Drilling) represents the first complete branch of the hierarchical tree of the classification scheme[2].

The investigation was triggered by our observation that applying baseline Machine Translation (MT) tools, such as Google Translate, to terms used in GICS produces substantially different terms in target languages than provided by the corresponding languages in GICS. For example, only a partial Spanish translation was obtained for the German compound ellipsis "Eigentums- und Unfallversicherungen", resulting in "Propiedad y accidente", whereas the correct translation should be "Seguro de Propiedad y Accidente" (Property and casualty insurance).

As regards structure, related work will be presented in section 2. Preprocessing steps and corrective patterns for the purpose at hand will be discussed in section 3. Deriving subcomponents of ontology labels for multilingual alignment will be the focus of section 4. Finally, the resulting ontology will be lexicalized by means of *lemon* [7] prior to concluding remarks.

2 Related Work

Research in various areas such as multilingual ontology acquisition [6], cross-lingual ontology mapping [11], ontology lexicalization [7], linguistic enrichment of labels [8], ontology engineering from text [9], and ontology localization [10]

[1] `http://www.standardandpoors.com/indices/gics/en/us`

[2] The definition associated with the leaf concept ID 10101010 is "Drilling contractors or owners of drilling rigs that contract their services for drilling wells."

can be observed. All of these approaches highlight the importance of ontologies labeled in different languages and techniques of acquiring them. While [11] seems to be the closest approach to our investigation, the major difference lies in the fact that [11] (and in fact also [15]) addresses only the language data included in RDF URI reference statements. Consequently, they are not concerned with natural language processing of (possibly lengthy) multilingual natural language strings, but only with finding equivalents of reference expressions in various lexical resources.

Current and future results of our work might best be compared to state-of-the-art research in the field of lexico-syntactic patterns, which are part of ontology design patterns[3] and mostly used for learning ontologies from natural language text (e.g. [5]). For this purpose and for the approach we apply to the analysis of the content of ontology labels, many different linguistic processes, such as tokenization, lemmatization, shallow parsing are used, also often combined with statistical machine learning techniques to learn ontologies from large sets of documents, e.g. Text2Onto [4]. The major problem of such patterns is low precision and over-generalization, which [3] try to overcome by restricting their main approach to three sets of patterns.

The creation of ontologies from text (e.g. [12, 2]) or other resources such as thesauri (e.g. [1]) and taxonomies has been a thriving research topic as of late. However, the use of multilingual information as a means of coherence and consistency check of ontology labels calls for further investigation. Our work seems to open the possibility to offer better proposals for the use of more consistent terminology in labels associated with ontology elements in a cross-lingual setting.

3 Initial Processing Steps and Cross-Lingual Corrective Patterns

We concentrate in this experiment on multilingual aspects in the GICS ontology we derived from the original taxonomy, having in mind the potential for an improved translation base for terms in this domain and for Information Extraction in documents describing among others activities of companies. Initially, we focused on labels in the three languages English, German, and Spanish, but have already experimented with Russian labels.

To remedy the deficient translatability of GICS labels, we investigated the transformation of the surface realization of the contained terms. In order to achieve a better readability of the ontology by engineers and users and better prepare labels to automatic processing, we transform non-lexical symbols to lexical correspondents, apply lexico-syntactic patterns to resolve compound ellipses, and complement labels based on constituency discrepancies across languages, i.e., missing constituents in one or more languages.

Replacing non-lexical items by their lexical correspondents refers to punctuation and ampersands. Duplicate occurrences of punctuation such as ",." are

[3] `http://ontologydesignpatterns.org`

corrected. Ampersands occur 159 times in the English taxonomy, the coordination word "and" not being used at all, while the German version features 117 occurrences and the Spanish only uses the coordination marker "y". The ampersand character serves to represent coordination, but automated linguistic decomposition of terms containing ampersands is not supported by off-the shelf NLP tools. As a rather straight-forward step the ampersand was replaced by "and" and "und" (DE).

At a more complex level we transform so called compound ellipses in GICS labels in fully lexicalized strings. Elliptical compounds represent the outcome of a deletion process of identical constituents in either the right or the left part of the coordination. For instance, the hyphenated German compound "Erdöl- und Erdgasförderung" (Oil and Gas Drilling) is transformed to "Erdölförderung und Erdgasförderung" (Oil Drilling and Gas Drilling). This transformation is not trivial as it requires both the analysis of the compounds and the resolution of the ellipsis, attaching the constituent "Förderung" to "Erdöl" in the example above. This process necessitated the use and adaptation of a morphological analysis component and the generation of ellipsis grammars, which are both implemented in the NooJ[4] finite state framework. Examples of the lexico-syntactic patterns implemented in NooJ are provided below.

[Examples of Resolution Patterns of Elliptical Coordinations]

```
DE: <NN1>hyphen und <NN2+NN3> resolved to <NN1+NN3> und <NN2+NN3>
EN: <NN1> and <NN2> <NN3>  resolved to <NN1> <NN3> and <NN2> <NN3>
ES: <NN1> <ADJA1> y <ADJA2> resolved to <NN1> <ADJA1> y <NN1> <ADJA2>

DE: <NN1+NN2> und hyphen<NN3> resolved to <NN1+NN2> und <NN1+NN3>
EN: <NN1> <NN2> and <NN3>  resolved to <NN1> <NN2> and <NN1> <NN3>
ES: <NN1> y <NN2> de <NN3> resolved to <NN1> de <NN3> y <NN2> de <NN3>
```

The presence of the German hyphen compound triggers the resolution of ellipses into coordinated structures in labels for other languages attached to the same concept. For instance, the German example above triggers the transformation of the English label "Oil and Gas Drilling" to "Oil Drilling and Gas Drilling" and of the Spanish label "Perforación de Pozos Petrolíferos y Gasíferos" to "Perforación de Pozos Petrolíferos y Perforación de Pozos Gasíferos". The resolution not only concerns single nouns, but also nominal phrases, e.g. "Perforación de Pozos", and adjectival phrases. As our algorithm requires the presence of a German hyphen, terms such as "Commercial Services and Supplies" (related to the German "Gewerbliche Dienste und Betriebsstoffe") are not resolved and are also not supposed to be resolved. All definitions attached to GICS terms confirm our approach to ellipsis resolution. Further examples of resolution in all three languages are as follows.

[Annotation Results of NooJ Processing applied to German, English and Spanish]

[4] http://www.nooj4nlp.net/pages/nooj.html

```
<EL TYPE="Energiezubehör#und#Energiedienst">Energiezubehör und -dienste</EL>
<ELLLL TYPE="Grosshandel#und#Einzelhandel">Gross- und Einzelhandel</ELLLL>

<EL TYPE="Energy#Equipment#and#Energy#Services">Energy Equipment and
Services</EL>
<EFOURD TYPE="Oil#:#Exploration#and#Oil#:#Production#and#Gas#:#Exploration#and
#Gas#:#Production">Oil and Gas Exploration and Production</EFOURD>

<EL TYPE="Equipos#de#Energía#y#Servicios#de#Energía">Equipos y Servicios de
 Energía</EL>.
<ELLL TYPE="Productos#Madereros#y#Productos#Papeleros">Productos Madereros y
Papeleros</ELLL>
```

At times the authors of the industry classification apply a colon to structure terms, such as *Metalle & Bergbau: Diverse* (Diversified Metals and Mining). Frequently, these constructs can only be resolved using prepositions instead of compounding, because terms such as *Heiwerkerausrüstungseinzelhandel* (Home Improvement Retail) do not exist. Structures using colons could only be observed in German labels of GICS.

As a final preprocessing step we evaluated complementing labels on the basis of a cross-lingual comparison. The German "Integrierte Erdöl- und Erdgasbetriebe" lacks any equivalent of "betrieb" (company) in the English or Spanish version. Despite the fact that the taxonomy is about business activities, the word company does virtually not occur in the English or Spanish designations of concepts, only in definitions. For the sake of completeness, we decided to complement the English and Spanish label with the equivalent of the missing term taken from sibling concepts in the same sector or definitions. In this case, we add "companies" and "empresas" on the basis of the assumption that multilingual labels associated with concepts should, where feasible, have the same amount and quality of information.

The presented algorithm ports all terms to a shared surface realization and depicts the different but aligned language specific realizations. While the patterns for resolving general ellipsis can be applied to other sources, such as the Industry Classification Benchmark (ICB)[5], the second case of terms separated by colon seems to be specific to GICS. Currently the algorithm has been implemented for the indicated languages, however, we have performed experiments with their utilization for other not closely related languages, such as Russian. Many lexico-syntactic patterns can be applied directly to the Russian designations, such as the compound "Хранение и транспортировка нефти и газа" (Storage and Transportation of Oil and Gas) can be resolved to "Хранение нефти и транспортировка нефти и Хранение газа и транспортировка газа" (Oil Storage and Oil Transportation and Gas Storage and Gas Transportation).

The representation of the fact that we modified the original terms (or labels) remains to be an issue. Indicating the modification is important to the authors of the taxonomy as well as people analyzing data. As a tentative step, for this

[5] http://www.icbenchmark.com/

purpose we have introduced the annotation property "preprocessed" to clarify that we have adapted the original content of labels and definitions.

4 Multilingual Alignment and Sub-Term Structures

Performing initial preprocessing steps facilitates the multilingual alignment of terms and components of terms. For the purpose of multilingual alignment, we have extensively analyzed and utilized existing hierarchical relations and definitions. In a second step we create relations to indicate sub-term relations in the actual ontology. By creating an additional terminological resource, we derive a second subsumption hierarchy focusing on sub-term relations, which is supposed to facilitate Information Extraction based on the ontology we created.

4.1 Term Alignment

Within the taxonomic structure of GICS we are able to establish relations between (sub)terms along the line of class hierarchies. GICS is structured along four major meta-categories in sector, industry group, industry, and sub-industry. Terms used in a super-class can thus be used for comparing a term in one language with the terms of other languages. Not only the line hierarchy is interesting for us but equally siblings in the hierarchy provide vital information.

Lexica and lexical resources created in the initial processing are now utilized to create multilingual alignments of the terminology contained in the taxonomy. We utilize lemmas of the normalized labels to facilitate the multilingual alignment as represented by the NooJ output illustrated below.

[Example of NooJ Annotation Result]

```
<TYPE="Integrierte#Erdoelbetriebe#und#Integrierte#Erdgasbetriebe">
Integrierte Erdoel- und Erdgasbetriebe</>
```

The associated lexical information in NooJ tells us in this case that "Integrierte" is the adjectival form derived from the verb "integrieren" (to integrate). The lemma of the head of the compound noun "betriebe" being then "Betrieb" (company). Thereby, we are able to establish term relations on the basis of the hierarchy, such as depicted below for the GICS class "101020".

[Example of Term Alignment]

```
"de"  => "Erdoel, Erdgas und nicht erneuerbare Brennstoffe",
"en"  => "Oil, Gas and Consumable Fuels",
"es"  => "Petroleo, Gas y Combustibles",

"trans"  =>
"Erdoel@de = Oil@en = Petroleo@es ::
 Erdgas@de = Gas@en = Gas@es ::
 Nicht erneuerbare Brennstoffe@de = Consumable Fuels@en = Combustibles@es"
```

Term pairs may vary strongly across different sectors within one classification. For instance, "Leisure products" equals "Freizeitartikel" in German, while "Agricultural Products" corresponds to "Landwirtschaftliche Produkte". Once "product" is aligned with "Artikel", in a different sector it maps to "Produkte". Nevertheless, this fact does not hamper automating the alignment process, which has been done on the basis of a Java tool, porting the preprocessed labels to the subsumption hierarchy of the ontology. At times, this initial alignment can lead to multiple mappings of terms depicted in Fig. 1.

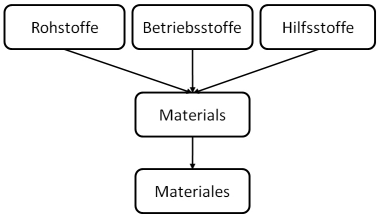

Fig. 1. Different Conceptualization of Cross-Lingual Designations

The interesting point about the example in Fig. 1 is the different conceptualization across languages. The German multi-word term corresponds to the single word expression "Material" in English and Spanish, which constitutes a challenge for cross-lingual alignment as there seems to be no equivalent for the three German expressions in the other languages.

In such cases other term pairs within the same sector are analyzed as regards re-occurrence of terms. If no equivalence can be detected, the definition has to be searched. Should the terms be only contained in one label, then additional resources, such as bilingual dictionaries or other multilingual industry classifications, might be consulted. However, in other cases clear misalignments occur, such as "Betriebsstoffe" in German being aligned to "Professional Services" (en) and "Servicios Profesionales" (es). As the same designation is part of another sub-industry in the sector, the incorrect alignment can be corrected on the basis of the existing correct alignment to "Professionelle Dienste". The definition in each language further confirms this alignment.

Our special focus is on terminal nodes in the original taxonomy as they contain detailed definitions, which further facilitates the cross-lingual alignment and validation of alignment correctness. As a tentative approach, we use lexico-syntactic patterns again to extract some basic information contained in definitions, exemplified by one pattern in German below. The extracted information as well as manually derived alignments from definitions are both used to validate the previously described alignments of designations of taxonomic concepts.

[Pattern for Extraction of Information in Definitions]

```
German:
<NP1>, die sich mit <NP2>, <NP3>und OR oder <NP4>
von <NP5> tätig sind OR beschäftigen.
Definition "Pharmazeutika": "Unternehmen, die in der Erforschung,
Entwicklung oder Herstellung von Pharmazeutika tätig sind."
```

Definitions provide further information to facilitate the construction of proper terms and term alignments. <NP1> represents a synonym of the word company, e.g. manufacturer, producer, provider, whereas the other noun phrases relate to business activities. One Spanish example is the industry of *Transportes*, which has the industry group *Transporto Aereo* and sub-industry *Lineas Aereas* referring to the former term explicitly in its definition.

Analyzing siblings creates relations that would otherwise not be evident. For instance, "Building Products" might not be related to "Aerospace and Defense" in any other domain. Within this sector, however, they are related as it regards the manufacturing of aerospace and defense equipment. The extracted term pairs of the definitions allow us to add these additional information to the label to strive towards completeness of information.

Terms aligned in this section are represented in the GICS ontology as annotation properties with the respective `xml:lang` property. Initial preprocessing and the correct alignment of terms serve to improve the overall quality of the natural language representation of the ontology. The alignment of terms equally helps to reveal inconsistencies or in other words improve the consistency of ontology labels.

4.2 Sub-Term Relations

At this point our ontology consists of five main classes according to the taxonomic structure, the four meta-levels and an additional class "Company". The latter features a "hasBusinessActivitiy" object property to the main class "Sub-Industry" so that upon instantiating a company various activities can be added. In addition, all taxonomic categories have a `subClass` relationship to the respective meta-category.

Creating sub-term relations introduces an additional structure not originally part of the GICS taxonomy, which is why we have decided to create an additional OWL-DL resource dedicated to terminology and terminological relations. For "isSubTermOf" relations it might be worth considering a transitive characteristic, that is: "P(x,y) and P(y,z) implies P(x,z)"[6], so each term y isSubTermOf x, z isSubTermOf y, which implies that z isSubTermOf x. This allows us to state that "Trucks" is a subterm of "Heavy Trucks" and at the same time of "Farm Machinery and Heavy Trucks". This type of decomposition abides by the terminological principles presented in ISO704:2009.

[6] `http://www.w3.org/TR/2004/REC-owl-guide-20040210/#PropertyCharacteristics`

In order to account for the terminological relations and levels, pseudo-categories, i.e., categories not originally part of the taxonomy and generated for terminological reasons, have to be introduced to the original hierarchy. This is due to the fact that terminological relations focus on hypernymic, meronymic relations. For example, the subcategories of *Energy* all refer to either Energy, Oil, Gas, or Consumable Fuels, all of which have to be introduced to the terminological structure.

The decomposition of e.g. "Oil Equipment and Gas Equipment and Oil Services and Gas Services" centers around the constituent and divides the term at the second "and". Accordingly, the definition of sub-industries has to be adapted to the changed concept and added to the terminological entry. Information extracted from definitions in the previous step are added to the terminology in order to enlarge the contained vocabulary.

A terminological representation of these natural language labels of an ontology provides a highly beneficial overview of contained terms, their sub-terms and relations between them. This facilitates duplicity and consistency evaluations of labels. In combination with part of speech, morphological, and syntactic information represented in *lemon*, there are various application scenarios from facilitating the creation of new labels to machine translation.

5 Lexicalizing Ontology Labels

Several approaches and models seek to provide a lexicon-ontology interface to reduce the complexity of the ontology, while at the same time providing full lexical information on the natural language representation of ontologies.

The *lemon* model [7] was developed within the Monnet project[7] and represents textual and linguistic information contained in ontologies as external RDF resource and establishes semantics by means of relating entries to the ontology, i.e., the relation represents a means to disambiguate words. It adapts the main principles of the Lexical Markup Framework (LMF) standardized in ISO 24613 and unites it with the core features of *LexInfo* in order to elaborate a specific ontology-lexicon model. Lexicon objects describe syntactic and morpho-syntactic properties, which are related to entities of the ontology via sense objects. Subsequent to applying state labels to the entry, i.e., preferred, alternative, hidden reference, the lexical sense links to the lexical entry, which might be decomposed to its individual elements.

Lexicons based on *lemon* can be created automatically by means of the *lemon* generator[8]. The following lexicon was created on the basis of the seed ontology, without any preprocessing and term alignment. As can be seen, decomposition of the term "Energy Equipment & Services" fails due to the ampersand and the ellipsis.

[7] `http://www.monnet-project.eu`
[8] `http://monnetproject.deri.ie/lemonsource/`

[*lemon* decomposition of "Energy Equipment & Services"]

```
<lemon:decomposition xmlns:ns0=
"http://www.w3.org/1999/02/22-rdf-syntax-ns#" ns0:parseType="Collection">
     <lemon:Component  rdf:about="unknown:/GICS__en/
               Energy%2BEquipment%2B%26%2BServices#comp">
               <lemon:element rdf:resource="unknown:/GICS__en/Energy"/>
     </lemon:Component>
     <lemon:Component rdf:about="unknown:/GICS__en/
               Energy%2BEquipment%2B%26%2BServices#comp2">
               <lemon:element rdf:resource="unknown:/GICS__en/Equipment"/>
     </lemon:Component>
     <lemon:Component rdf:about="unknown:/GICS__en/
              Energy%2BEquipment%2B%26%2BServices#comp3">
             <lemon:element rdf:resource="unknown:/GICS__en/Services"/>
     </lemon:Component>
</lemon:decomposition>
```

The application of off-the-shelf NLP tools to labels in fact negatively influences the efficiency of an automated *lemon* based lexicalization process of labels, as most commonly used tools are not in the position to handle such types of (mainly nominal) ellipsis. Considering the fact that ontology labels to a large extend only consist of nouns and noun compounds, the issue is a vital one. We apply the process of lexicalization to the annotation property `rdfs:label` available in all languages covered in the GICS ontology, namely German, English, Spanish. For this purpose we use *lemon* for the representation of linguistic information added to these labels and linking to the original ontology elements.

Lexicalization supports the decomposition of terms into sub-terms, that is it facilitates the application of patterns to detect cross-lingual alignments at the level of components of terms/labels. The linguistic information in the *lemon* representation is being used for consolidation. However, we consider the decomposition of terms to be part of the terminological level, thus, introducing the terminological resource for GICS in section 4. The example below shows the encoding of constituency and part-of-speech information subsequent to our initial preprocessing and term alignment process.

[Constituency and Part-Of-Speech Information of "Energy Equipment and Energy Services" in *lemon*]

```
<lemon:entry>
     <lemon:LexicalEntry rdf:about="unknown:/lexicon__en/Energy+Equipment+and+Energy+Services">
        <lemon:sense>
          <lemon:LexicalSense rdf:about="unknown:/lexicon__en/Energy%2BEquipment%2Band%
             2BEnergy%2BServices#sense">
            <lemon:reference rdf:resource="http://www.semanticweb.org/ontologies/2012/8/GICS.owl#GICS101010"/>
          </lemon:LexicalSense>
        </lemon:sense>
        <lemon:canonicalForm>
          <lemon:Form rdf:about="unknown:/lexicon__en/Energy+Equipment+and+Energy+Services#form">
            <lemon:writtenRep xml:lang="en">Energy Equipment and Energy Services</lemon:writtenRep>
          </lemon:Form>
        </lemon:canonicalForm>
        <lemon:phraseRoot>
          ...
          <lemon:constituent rdf:resource="http://monnetproject.deri.ie/tags/penn/node/NN"/>
          ...
          <lemon:constituent rdf:resource="http://monnetproject.deri.ie/tags/penn/node/NNS"/>
          ...
          <lemon:constituent rdf:resource="http://monnetproject.deri.ie/tags/penn/node/NP"/>
```

```
            ...
            <lemon:constituent rdf:resource="http://monnetproject.deri.ie/tags/penn/node/CC"/>
            ...
            <lemon:constituent rdf:resource="http://monnetproject.deri.ie/tags/penn/node/NN"/>
            ...
            <lemon:constituent rdf:resource="http://monnetproject.deri.ie/tags/penn/node/NP"/>
            ...
            <lemon:constituent rdf:resource="http://monnetproject.deri.ie/tags/penn/node/NP"/>
...
</lemon:entry>
```

Due to space constraints the example only provides an English version, however, the same improved results can be observed in German and Spanish. The above example provides that *lemon* was in the position to decompose the term and provide part-of-speech information, using the Penn Treebank Notation. The lexical sense contains the link to the ontology and the original label as "writtenRep", followed by information on individual elements of the term. This use case is supposed to show that that such type of preprocessing and term alignment has beneficial effects on ontology labels.

6 Concluding Remarks and Future Work

We have preprocessed the labels of an ontology we derived from the GICS taxonomy, for the time being in English, German, and Spanish. We showed a pattern-based approach to resolving compound ellipses, which can be generalized across resources, such as the Industry Classification Benchmark (ICB). Thereby, we created terms initially not contained in the resource and thus, inaccessible to ontology-based tasks, such as Information Extraction. We aligned the terms across all three languages. Terms contained in definitions were extracted and additionally aligned to increase the overall quality and validate existing alignments. Furthermore, the normalized and aligned terms were included in a terminological resource in OWL-DL to provide explicit sub-term relations and decompose complex, long labels. Lexicalizing the derived ontology with its processed labels as opposed to the initial ontology served to exemplify the usefulness of such (pre)processing of labels.

As regards future work, we are currently investigating the applicability of our pattern-based approach to other language families than Romance languages. One further approach that might be interesting is the automation of the creation of a terminological resource for the ontology, similar to the idea of the *lemon* generator.

References

1. Kless, D., Jansen, L., Lindenthal, J., Wiebensohn, J.:A Method of Re-Engineering a Thesaurus into an Ontology. In: Donelly, M., Guizzardi, G. (eds): Formal Ontology in Information Systems - Proceedings of the Seventh International Conference (FOIS 2012), pp.133–146. IOS Press, Amsterdam (2012)
2. Serra, I., Girardi, R.: A Process for Extracting Non-Taxonomic Relationships of Ontologies from Text. Intelligent Information Management 3, 119–124 (2011)

3. Maynard, D. F. A., Peters, W.: Using lexicosyntactic Ontology Design Patterns for Ontology Creation and Population. In Proceecdings of the Workshop on Ontology Patterns (2009)
4. Cimiano, P., Voelker, J.: Text2Onto - A Framework for Ontology Learning and Data-driven Change Discovery. In: Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB), Alicante, Spain (2005)
5. Klaussner, C., Zhekova, D.: Lexico-Syntactic Patterns for Automatic Ontology Building. In: Proceedings of the International Conference Recent Advances in Natural Language Processing (2011)
6. Nichols, E., Bond, F., Tanaka, T., Fujita, S., Flickinger, D.: Multilingual Ontology Acquisition from Multiple MRDs. In Proceedings of the 2nd Workshop on Ontology Learning and Population, pp. 10–17 (2006)
7. Buitelaar, P., Cimiano, P., McCrae, J., Montlie-Ponsoda, E., Declerck, T.: Ontology Lexicalization: The *lemon* Perspective. In: Slodzian, M., Valette, M., Aussenac-Gilles, N., Condamines, A., Hernandez, N., Rothenburger, B. (eds.): Workshop Proceedings of the 9th International Conference on Terminology and Artificial Intelligence, pp- 33–36, Paris, France, INALCO, Paris (2011)
8. Declerck, T., Lendvai P.: Towards a standardized linguistic annotation of the textual content of labels in knowledge representation systems. In: LREC 2010- The seventh international conference on Language Resources and Evaluation. Interna- tional Conference on Language Resources and Evaluation (LREC-10), Malta (2010)
9. Aussenac-Gilles, N., Szulman, S., Despres, S.: The Terminae Method and Platform for Ontology Engineering from Texts. In Proceedings of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge. IOS Press, pp. 199–223, (2008)
10. Mejía, M.E., Montiel-Ponsoda, E., Aguado de Cea, G., Gómez-Pérez, A.: Ontology Localization. In: Suárez-Figueroa, M.C.,Gómez-Pérez, A., Motta, E., Gangemi, A. (eds): Ontology Engineering in a Networked World. pp. 171–191, Springer Berlin Heidelberg (2012)
11. Fu, B., Brennan, R., O'Sullivan, D.: Using Pseudo Feedback to Improve Cross-Lingual Ontology Mapping. In: Proceedings of the 8th Extended Semantic Web Conference (ESWC 2011), LNCS 6643, pp. 336–351 (2011)
12. de Cea, G.A., Gómez-Pérez, A., Ponsoda, E.M., Suárez-Figueroa, M.C.: Natural Language-Based Approach for Helping in the Reuse of Ontology Design Patterns. In: Proceedings of the 16th International Conference on Knowledge Engineering and Knowledge Management Knowledge Patterns (EKAW 2008), Acitrezza, Italy (2008)
13. Suárez-Figueroa, M.C., Gómez-Pérez, A., Fernández-López, M.: The NeOn Methodology for Ontology Engineering. In: Suárez-Figueroa, M.C.,Gómez-Pérez, A., Motta, E., Gangemi, A. (eds): Ontology Engineering in a Networked World. pp. 9–34, Springer Berlin Heidelberg (2012)
14. Cimiano P., Buitelaar P., McCrae J., Sintek M.: LexInfo: A declarative model for the lexiconontology interface. Journal of Web Semanics, Vol. 9, No. 1, pp. 29–51 (2011)
15. Vertan, C., v.Hahn, W. Challenges fort he Multilingual Semantic Web. In Proceedings of the International Workshop on Semantic web Technologies for Machine Translation, in conjunction with MT-Summit X, Phuket, Thailand (2005)

Experiences with Multilingual Modeling in the Development of the International Classification of Traditional Medicine Ontology

Csongor Nyulas, Tania Tudorache, Samson Tu, Mark A. Musen

Stanford Center for Biomedical Informatics Research, Stanford University, US
{nyulas, tudorache, swt, musen}@stanford.edu

Abstract. The World Health Organization (WHO) in collaboration with several international stakeholders have started recently the work on the International Classification of Traditional Medicine (ICTM), which will provide a standardized system for encoding and collecting health statistics data related to Traditional Medicine practice throughout the world. ICTM is represented in OWL, and is developed by Traditional Medicine experts in a collaborative Semantic Web platform, called iCAT-TM. The content of ICTM is developed simultaneously in four languages (English, Chinese, Japanese and Korean). In this paper, we describe how we modeled the multilingual content, the Web platform used for editing, and some of the challenges we have encountered related to the multilingual aspects of the model and use of the platform.

1 The International Classification of Traditional Medicine (ICTM)

The World Health Organization (WHO) in collaboration with a large group of international stakeholders is developing the International Classification of Traditional Medicine (ICTM).[1] ICTM will provide a standardized international system for classifying Traditional Medicine (TM) related health concepts, such as disorder names, disease patterns, signs and symptoms, causal factors, and interventions [9]. One of the goals of the project is to be able to unify the data collection and monitoring for Traditional Medicine systems with those of the conventional (i.e., "Western") medicine, which will be realized by integrating a relevant part of ICTM as Chapter 23 of the 11th revision of the International Classification of Diseases (ICD).[2] ICD is an essential classification used in the United Nation countries for compiling basic health statistics, billing, and clinical documentation [8].

The content of ICTM is based on classifications of Traditional Medicine from three countries, China, Japan and Korea. Even if these classifications have a common root, they have diverged significantly over the years. The role of ICTM is to harmonize these different efforts and come to a consensus classification that can be used in health systems around the world.

With the information age revolution, WHO has changed significantly the way they build classifications. To make them ready for electronic health records and enable easy

[1] https://sites.google.com/site/whoictm/
[2] http://www.who.int/classifications/icd11/browse/f/en

cross-linking between them, the classifications have now a formal underpinning. ICTM, similarly to ICD-11, is represented as an OWL ontology and is developed using Semantic Web technologies.

Given the international nature of ICTM, tackling the multilinguality problem is one of the main challenges in the project. Domain experts from the three countries and the project coordinators in Geneva, Switzerland, are developing the content of ICTM simultaneously in four languages: English, Chinese, Japanese, and Korean. Our group has provided the ontology modeling support and the Web platform infrastructure used for editing ICTM. In this paper, we describe our experiences in supporting multiple languages in the ICTM ontology, including the model and tooling, and the challenges we encountered.

The rest of the paper is organized as follows: Section 2 describes the related work, in Section 3, we describe how we modeled the multilingual content in ICTM, Section 4 presents the collaborative Semantic Web platform used by the domain experts to edit ICTM, and finally, Section 5 presents the challenges and some lessons learned in the project, and gives an overview of the future work.

2 Related Work

As the Semantic Web matures, there is an increasing body of research on localizing ontologies. For example, the SKOS-XL extension [1] treats labels as first order resources, thus enabling the definition of explicit links between labels associated to the same concept. Montiel-Ponsoda *et al.* [4] try to overcome some of the limitations of the SKOS-XL representation and propose a module for lemon [3] that supports different types of translation relations and metadata, such as provenance and reliability scores. Extensive work on ontology localization [2] has also been done in the NEON project[3] that proposes guidelines and a tool to support this process.

Silva *et al.* present conceptME [5], a collaboration framework that supports ontology localization starting early in the conceptualization phase. Providing terminological support so early in the development process proved to enhance the conceptualization of the domain. conceptME has also support for sharing conceptual models, for content negotiation and discussion.

In this work we did not use any of the related approaches, as one of the main requirements in the project (see Section 3) was to use and/or extend the ICD-11 ontology to ensure that these two ontologies will be easily integrated at a later stage. We plan to investigate the related approaches (such as SKOS-XL and the extensions to lemon) to see if they would fit the requirements for ICTM, and if so, we will refactor our ontology accordingly.

3 Multilingual Modeling in ICTM

As we mentioned before, one of the main requirements for ICTM is that it should follow similar modeling patterns to the ICD-11 ontology [6], so that these two can be easily integrated.[4] In addition, all ICTM textual content should be available in four languages: English, Chinese, Japanese and Korean, which Traditional Medicine experts

[3] `http://www.neon-project.org`

[4] As we mentioned before, part of ICTM will be available as a separate chapter in ICD-11.

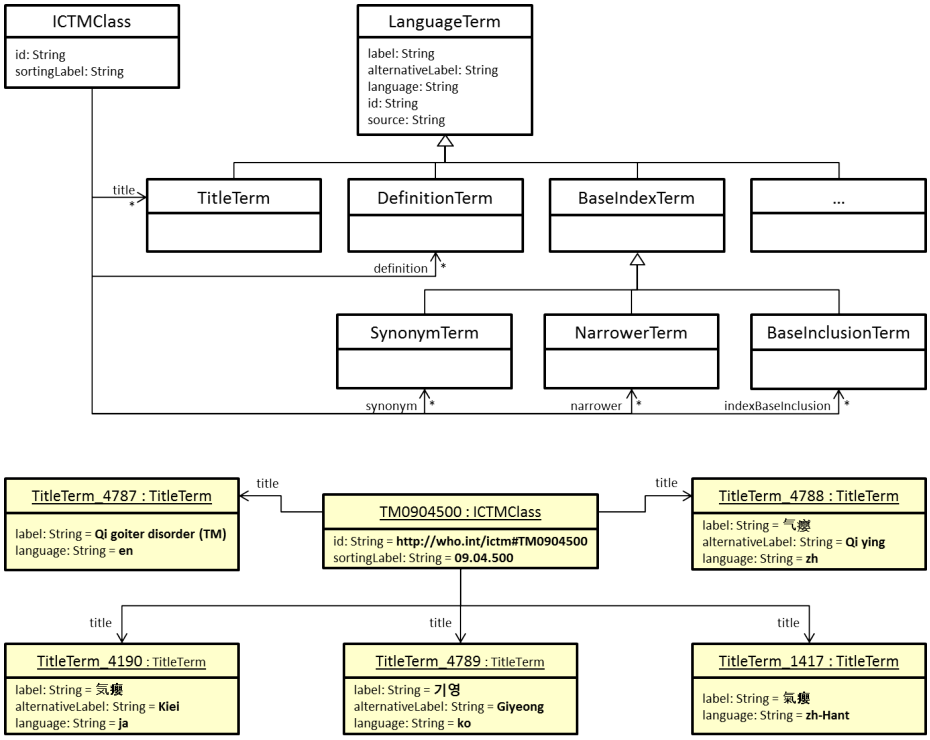

Fig. 1. Excerpt from the ICTM ontology. Language terms are modeled as instances of the reified class *LanguageTerm*. Subclasses of *LanguageTerm* represent different linguistic terms (title, definition, synonym, and so on). The subclasses may have additional properties that represent different metadata of the language term. The class level is shown in boxes with white background. We also show an example instantiation for the title terms for the *Qi goiter disorder* disease class using the boxes with darker background.

from different countries will input during development. A further requirement, which came later in the project, was to support transliteration of titles, i.e., converting the Chinese, Japanese and Korean scripts into Latin script. For example, a common transliteration for converting Chinese characters into Latin script is Pinyin. Figures 1 and 2 show the transliteration of the simplified Chinese disease title 气瘿 (meaning, *Qi goiter disorder*) into Pinyin as *Qi ying*. Other metadata will be attached to the label of a term into a specific language, such as the source of the label (e.g., the Traditional Medicine classification where the label originates from), and an internal id that is used by other WHO software.

We modeled ICTM in OWL 1.0. We used a reified class, *LanguageTerm*, to represent all linguistic terms in the ontology. We have created a taxonomy of language terms as subclasses of the *LanguageTerm* class, as some of the term types have additional properties attached to them. For example, the *SynonymTerm* has additional properties that describe if and how it will be included in an electronic index for the classification.

In the current version of the model, there are eight subclasses of *LanguageTerm* that represent among others, the title, the fully specified title, the definition, other external definitions, the synonyms, and so on. The actual value for a language term is an instance of a subclass of *LanguageTerm*.

Figure 1 shows an excerpt of the class level modeling for language terms and how different properties of a disease class (title, definition, synonym, etc.) have been reified. The figure also shows an example for modeling the five title terms for the *Qi goiter disorder* disease class. The *Qi goiter disorder* class has a property *title* that has as values five instances of the class *TitleTerm* that correspond to the titles in 5 languages (English, Japanese, Korean, simplified Chinese and traditional Chinese). Some *TitleTerm* instances (e.g., *TitleTerm_4788* for simplified Chinese) has in addition to the *label* and *language* properties, also another property, *alternativeLabel*, to represent the transliteration of the Chinese script to Latin characters. Other language terms (not shown in the figure), such as the *ExternalDefinitionTerm*—used to reference textual definitions from external resources — have additional properties that specify the source of the definition in greater detail (e.g., the ontology name, the IRI of the source ontology entity, the URL for the source ontology, etc.).

As there might be confusion about the difference between synonyms and transliterations, we would like to clarify this issue. A synonym is a term that has a similar meaning to a another term (in our case, the title term). In ICTM, as in ICD-11, synonyms are also used to store alternative titles for a disease, that are either found in scientific literature, or have minor linguistic variations, or are used in the colloquial language (e.g, the synonym for *Roseola infantum* is *Sixth disease*). The synonyms apply for terms in the same language (e.g., an English title may have other English synonyms). A transliteration, on the other hand, represents exactly the same term in the same language, but in a different script. A term may have several transliterations (Korean has 4 different transliterations).

4 The iCAT-TM Platform

Traditional Medicine experts around the world are editing ICTM using the collaborative iCAT-TM Web platform. iCAT-TM is a customization of the generic WebProtégé ontology editor [7]. The user interface of iCAT-TM is tailored for domain experts, who are not knowledgeable about ontologies or knowledge representation. iCAT-TM presents a form-based interface shown in Figure 2 that is is less intimidating for the experts than a generic ontology editor would be. The experts can edit the class taxonomy in the left panel of the *ICTM Content Tab*, and the class details, including the language terms, in the right panel.

iCAT-TM has many collaboration features inherited from WebProtégé, such as the support for simultaneous editing, change history of users' actions, and notes and discussions attached to any entity in the ontology.

We have created a generic widget that displays the content of reified individuals, and we have reused it for displaying and editing the different language terms. In Figure 2, the *ICTM Title* uses this widget to display the values of the *TitleTerm* individuals associated to the *title* property of a disease. A row in the widget table corresponds to one of the reified individual values, and the columns display the properties of the respective row individual value. The same widget is also used for displaying the short definition of a disease (the *transliteration* column has been hidden from view).

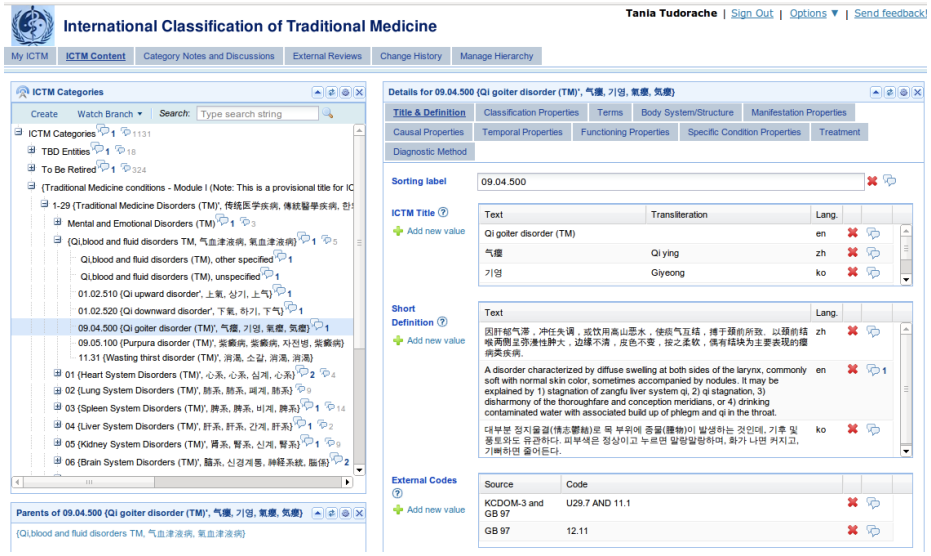

Fig. 2. The iCAT-TM platform is used by domain experts from the three countries to develop ICTM collaboratively on the Web. The panel on the left hand side shows the class tree, and the right hand side panel shows the details of the selected class (in this case, the *Qi goiter disorder*). The language terms for the title and short definition are also shown, as well as the transliteration for the title.

One "bonus" of using reified individual for language terms (which, as a consequence, have identity) is that we can attach notes and discussion threads to a particular individual. For example, in Figure 2, the second short definition of the disease has a comment attached to it (shown as the number *1* next to the comment icon on the second row). This feature enables domain experts to have focused discussions right in the context in which they are editing. The contextual discussions are particularly useful because each disease has several properties that need to be filled, in many cases by different experts, and the overview and management of notes and discussions is much easier.

5 Discussions and Future Work

The iCAT-TM has been in production use since February 2011 by 25 Traditional Medicine experts. As a result, ICTM contains now more than 1,500 classes, 15,000 reified terms, out of which, 10,000 are language terms. The users have created more than 60,000 changes in the ontology, and added more than 1,100 notes and discussions.

Since the beginning of the project, we have encountered several challenges related to the multilingual aspects in the modeling, tooling and use of the platform.

Modeling. ICTM was developed using OWL 1.0 to make it compatible with the ICD-11 ontology. For this reason, we had to use reified relations to model the language terms. Reified relations, even though they have the advantages described earlier, have several disadvantages, as well. First, the reified individuals clutter the domain ontol-

ogy, and increase its size significantly (in ICTM, almost all property values are reified). Second, these anonymous individuals are used in reasoning (as part of the domain ontology) and can slow it down significantly. We plan to overcome these limitations by upgrading the ontology to OWL 2.0 (ICD-11 will also upgrade), and rather than using reified individuals, we plan to use annotations on axioms. We plan to change the modeling in other aspects, too. For example, the transliterations are currently modeled as a multiple cardinality datatype property that take string literals as values. Even if we can now add more transliterations for the same label (e.g., Korean has four different transliterations), we cannot specify to which script or alphabet a transliteration belongs to. We plan to address this issue by using nested annotations on axioms in the OWL 2.0 modeling. Additionally, we plan to investigate if other approaches for ontology localizations, such as the ones we mentioned in the Related Work section, are suitable for ICTM. If these approaches fit the requirements, we will refactor the ontology to use a more standard approach. This undertaking will, however, require significant effort, as we need to also change the modeling of ICD-11, as well as migrate all existing content of two live production system (iCAT for ICD-11, and iCAT-TM for ICTM) to the new structure.

Tooling. We had to make sure that our tooling works well with international characters. While these are not an issue for the Web application per se (Web browsers can show pages in different encodings), we had to adjust our Lucene-based search mechanism to work properly with multiple languages. One hurdle for the domain experts in using iCAT-TM is that the user interface is presented in English, and many of them are not very comfortable with it. We plan to redesign the user interface to better follow the principles of internationalization, so that we can more easily provide language specific user interfaces. We do expect that this step will involve a significant re-design effort.

Use of the platform. We had several user related challenges that are not necessarily of technical nature. For example, when we started the project, we used (wrongly) the country codes to model the languages (*ch, jp*, and *kr*). Later, in the process, we changed the language codes to the correct ones from the ISO 639-1 (*en, zh, ja, ko*), however, some of the domain experts complained that the correct language codes are less intuitive to use. Also, when we started the project, we did not anticipate that some content will be entered in simplified Chinese, while other will be entered in traditional Chinese, which created some confusion with the users. As a solution, we added also the traditional Chinese language code (*zh-Hant*), so that at a later date the Chinese content can be easier curated and harmonized (it is expected that in the official distribution only simplified Chinese will be used). Another challenge is related to the communication among the domain experts, as most of them speak only their native language, and sometimes English, too. To improve the communication among the domain experts and the WHO coordinators in Geneva, we have introduced the transliteration. Another challenge related to the language barrier is that experts do not agree on the English translation for a term, and "invent" new English translations. This fact also makes the curation and verification of the entire classification content very challenging, because finding Traditional Medicine experts who understand all languages and can verify that the terms in different languages really mean the same thing, is very difficult.

As future work, we plan to upgrade ICTM to OWL 2.0 to overcome the modeling issues we described before. We will also create linkages between ICTM classes and ICD-11 classes that will put into correspondence Traditional Medicine disorders with "Western" diseases. As the project progresses, we will also provide a peer-reviewing mechanism, in which external domain experts will review different aspects of ICTM.

The iCAT-TM platform is currently in production use, and we expect that by 2015, when the ICD-11 major revision is planned to end, the ICD-11 Chapter 23, containing a part of ICTM, will be finalized as well. Even after 2015, ICTM will continue to be developed as an independent classification that will address the needs of the Traditional Medicine practices around the world.

Acknowledgments

We thank our WHO collaborators and the ICTM project members for developing the project requirements and for the fruitful collaboration. The work presented in this paper is partly supported by the NIGMS Grant 1R01GM086587.

References

1. SKOS Simple Knowledge Organization System eXtension for Labels (SKOS-XL) Namespace Document - HTML Variant. `http://www.w3.org/TR/skos-reference/skos-xl.html`. Last accessed: August, 2012.
2. M. Espinoza, E. Montiel-Ponsoda, and A. Gómez-Pérez. Ontology localization. In *Proceedings of the fifth international conference on Knowledge capture*, pages 33–40. ACM, 2009.
3. J. McCrae, G. Aguado-de Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gomez-Perez, J. Garcia, L. Hollink, E. Montiel-Ponsoda, D. Spohr, et al. Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*, 2011.
4. E. Montiel-Ponsoda, J. Gracia, G. Aguado-de Cea, and A. Gómez-Pérez. Representing translations on the semantic web. *MSW 2011*, page 25, 2011.
5. M. Silva, A. Soares, and R. Costa. Supporting collaboration in multilingual ontology specification: the conceptme approach. *TKE 2012*, page 27.
6. T. Tudorache, S. Falconer, C. Nyulas, N. Noy, and M. Musen. Will Semantic Web Technologies Work for the Development of ICD-11? In *The 9th Intl. Semantic Web Conference (ISWC 2010)*, pages 257–272. Springer, 2010.
7. T. Tudorache, C. Nyulas, N. Noy, and M. Musen. Webprotégé: A collaborative ontology editor and knowledge acquisition tool for the web. *Semantic Web Journal*, pages 1–11, 2012.
8. World Health Organization. International Classification of Diseases (ICD). `http://www.who.int/classifications/icd/`. Last accessed: August, 2012.
9. World Health Organization. Traditional Medicine in Health Information Systems: Integrating Traditional Medicine into the WHO Family of International Classifications. `https://sites.google.com/site/whoictm/home/ICTMProjectPlan.pdf`. Last accessed: August, 2012.

Hybridizing formal and linguistic semantics for the MSW

Aldo Gangemi[1, 2]

[1]Paris13-CNRS-Sorbonne Cité and [2]STLab, ISTC-CNR, Rome
aldo.gangemi@cnr.it

1. Semantics: a serendipitous chaos

The current uptake of "semantic technologies" requires an effort to design some interoperability for the representation practices among fields as diverse as knowledge representation and reasoning (KR), lexical semantics, information extraction, databases, (semantic) Web standards, Web 2.0 folksonomies, etc.

Multilingual linguistic elements, *ontologies*, and *semantics* are key components that are shared by those fields, but are approached in heterogeneous ways. Due to the enormous amount of legacy data and representation practices, we cannot count only on standardisation efforts to build useful applications. In the forthcoming Multilingual Semantic Web (MSW), we need to live with the "serendipitous chaos" that characterises knowledge (and linguistic) management and engineering.

Too rigorous requirements are not sustainable, as the history of the Semantic Web in the last ten years suggests: logical consistency cannot always be enforced, identity of entities is often questionable, data are not always reliable and usually incomplete, knowledge can take many forms, assignment of predicates to objects can be made for unpredictable reasons, and can change dynamically, the intended meaning of predicates cannot even be studied to a full extent, because any two persons can have various levels of competences, and different needs for their interaction with their environments, often entering a dialectic or even conflicting interaction. Even more importantly, data and content are rarely structured in a cognitively sound way, or in a way that is *relevant* to the humans or applications that use them [1].

For those reasons, we have requirements for an agile semantics that (1) overarches the different representation practices, (2) is able to deal with incompleteness and errors, but also (3) assumes cognitive relevance by default.

In everyday life, any sign that we use or perceive (the perception of a segment of the world, an image, a word, a sentence, a scientific handbook, a novel) is not typically interpreted as it is supposed to be according to an ontology, dictionary, or other quasi-normative resources, but as a function of what we can do with it, i.e. as a relevance function, also known as an *affordance* [2]. For MSW this is a very important assumption, because when we envisage applications that are cross-linguistic, they need to work at the level of cognitive relevance, not at that of single, decontextualized data or term equivalences.

A representation language that integrates ontologies and (multi-, cross-)linguistic data needs then to assume that a sign is interpreted (or produced) with an interaction context in mind. In addition, such representation language should be associated with the practices of accessing, reengineering, or refactoring data when used for a certain

adfa, p. 1, 2011.

purpose, e.g. with natural language processing methods, ontology-based data access, etc., including practices of multilingual corpora matching.

My position, which supports a preliminary sketch of FRASL (FRame ASsignment Language) in later sections, is that we need to define a minimal logical backbone (requirements (1)(2)), and to go back to the (relevance-based) cognitive foundations of KR, which was shared in the seventies (then lost) among AI, linguistics, and cognitive science researchers (requirement (3)), and revisit the way we design ontologies and data accordingly, in the MSW perspective.

2. A minimal model of semantic assignment

Inspired by [3][4][5], I assume folksonomies as used on typical Web2.0 applications as bearing the minimal semantic commitment for our problem. As Figure 1 summarizes, we can imagine a double nature of tagging/annotation on the Web, i.e. that *tags* are *assigned* (and *providing access*) to *resources*, so that the *label* used as face value of that tag expresses a *concept*. Also, a shared assumption on the Semantic Web (and annotation semantics in general) is that those concepts are *instantiated* by the annotated resources.

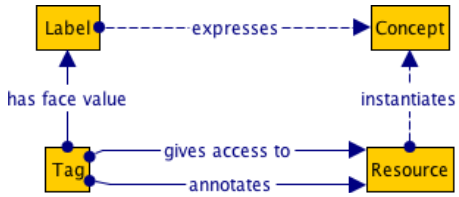

Figure 1: assignment operations and their semantic consequences. Dashed arrows denote the indirect nature of the semantics emerging from assignments.

Of course, there are big differences in labels taken from a folksonomy, extracted from a text, or defined in a formal ontology. The differences are mainly reflected in the way the concept is expected to be interpreted. For example, a label from a Web2.0 tagging action is simply interpreted from the combination of its bare label and the annotated resource(s). A label extracted from text is interpreted also with reference to the text itself, or other text/knowledge known as related to it. Finally, a label from the signature of a formal ontology is interpreted only (or mainly) with respect to its *formal semantics*.

However, despite the differences, the evolution of linked data and semantic applications show that, whatever additional constraints are given in a vocabulary or an ontology, the primary interpretation comes from the *intention of the tagger*, as one can notice from the wild usage of `owl:sameAs`, or the creative reuse of existing vocabularies.

Based on the cognitive semantics hypothesis, the intention of the tagger can be conceived as the relevance function applied in the tagging/annotation action. I call this action *assignment*. Assignments do not require any standpoint on the purely se-

mantic layer: the world of semantics is then accessory, and can be exploited for any added value it can provide besides the basic investigation of assignment actions.

This move frees up the possibility of a KR language that can deal with even purely geometrical accounts of meaning (e.g. from latent semantic analysis, social network analysis, clustering, multi-lingual corpora analysis, etc.), which only work on regularities (patterns) emerging from annotation practices, i.e. devoid of any high-level semantic standpoint.

A notable result is also that *formal* and *linguistic* semantics can be reconciled, provided that they are both grounded in assignments. For example, on one hand the formal interpretation of *hospital* is usually given as the class of 'all' hospitals, but in an assignment-based domain, the class of hospitals is the set of entities that are invariant under certain conditions deriving from compatibility of tagging operations by different agents and with an equivalence class of labels. On the other hand, the linguistic semantics of *hospital* will derive directly from the compatibility of tagging operations, eventually gathering the same grounding as the formal interpretation. An interesting consequence is that within empirically established assignment domains, we can use lexical concepts as formal classes, and vice-versa.

Moreover, my position is that concepts depend on the *relevance function* applied with the assignment. From the hypotheses, relevance functions activate real, fictionary, imagined, or simulated *action (*or more generally *situation) possibilities*. This is what notions like *frame, schema, script*, or *knowledge pattern* typically convey. Frame semantics in this perspective has been reconstructed in [whatsinaschema][towards][cahiers]. The consequence of this position is that whenever we extract or reuse a concept in an assignment scenario, that concept is either a frame (situation type, event type, etc.), or a role of a frame, or a type of a role from a frame. For example, assignment semantics assumes that the label *dog* has only sense in the context of a situation or action where a dog has a role, e.g. *barking* or *chasing*. Any multilingual treatment of *dog* will then need to cope with the contextual binding of that label.

Beaugrande [6] firstly defines "global patterns of knowledge" as a notion encompassing *frames*, *schemas*, *plans*, and *scripts*. Following him, as well as recent work in KR and the Semantic Web [7][1], we call this core notion *knowledge pattern*.

Knowledge patterns seemed appropriate in the seventies to create a positive cross-disciplinary research synergy. KR had a major role in this synergy. Description Logics were among the designs proposed, and for several reasons managed to be a major part of the development of the Semantic Web until nowadays. While DL have been very helpful in understanding the complexity problems behind automated reasoning on frame-like formal languages, they are rather poor when representing sorts like *frame*, *role*, *lexical unit*, *context*, *situation*, etc.

3. FRASL

The proposal that we briefly present here of a FRame ASsignment Language (FRASL), presented fully in [13], derives from previous work (e.g. [8]), but it stands alone in terms of practically covering the wide range of transformations and applications related to the ontology-lexicon interface. FRASL framework has several inspira-

tions, the most evident being Davidson's theory of events [9], Smith's *descriptions* [10], Construction Grammar [11], Discourse Representation Theory [12], etc.

The starting point of FRASL is the *Assignment* relation. An assignment is a semiotic action performed by some *Agent*, during either the production or the interpretation of a discourse fragment, called *Expression*, in an interaction between that agent and its *Environment*, in order to select a *Situation* from the environment.

Frame semantics tries to describe how situations are selected. *Frames* (or knowledge patterns) are situation types featuring roles that are filled by entities of a situation: in this way, situations emerge by filling the role structure of a frame. For example, in the **Cure** frame, a ***healer*** treats an ***affliction*** of a ***patient***, using some ***treatment*** (at some ***time***, ***place***, etc.). If an environment offers entities (e.g. a physician, a medical record, an injured person, and some medicaments) that fill the roles of that frame, we can recognize a *curing* situation within that environment.

In many cases, assignment operations do not provide extensive expressions; i.e. the frame **Cure** can be activated ("evoked") even by the picture of a hospital or a sufferer, the tags *healing* or *emergency* annotating a picture, or a sentence like: *he will undergo radiation treatment*.

A FRASL formula comprises components to represent the elements of assignment operations. For example, this is a template of FRASL components:

(1) **Scope**{"Expression" > (**frame**[***role***:entity(Type), ...]$_{\textit{situation}}$)}

Except scopes, any component can be empty. E.g. this template is almost empty:

(2) **Scope**{(**frame**[])}

For example, sentence (9):

(3) Mustafa said he decided to go alone to Socotra

can be represented as in formula (10):[1]

(4) **Sentence**{"Mustafa said" >
(**say**[***agent***:Mustafa(*x*:Person), ***time***:past(Time), ***sentence***:"he decided" >
(**decide**[***agent***:*x*, ***time***:past, ***sentence***:"to go alone to Socotra" >
(**go**[***agent***:*x*, ***location***:Socotra(Place), ***manner***:**alone** [***agent***:*x*]])])])}

The format of the predicates in (4) reflects that FRASL is a strongly-typed language: besides variables and named entities (individual constants), predicative constants can be sentence types, frames, roles, types, or modal modifiers.

In order to ground FRASL in a formal semantics, we need at least a translation to a many-sorted logic with proposition variables,[2] which gets a formal interpretation from

[1] See [13] for a detailed explanation of the FRASL notation.

[2] The following formula is semantically equivalent to (4): $\exists(x,y,t,z,g,w,p,a)$(say(x,y,t) ∧ agent(say,x) ∧ time(say,t) ∧ sentence(say,y) ∧ Person(x) ∧ x=*Mustafa* ∧ t=past ∧ y="he decided"∧ expresses(y,z) ∧ g="to go alone to Socotra" ∧ expresses (g,w) ∧ z=(decide(x,w,t)) ∧ w=(go($x,Socotra,a$)) ∧ agent($decide,x$) ∧ sentence($decide,w$) ∧ time($decide,t$) ∧ agent(go,x) ∧ location($go,Socotra$) ∧ Place(*Socotra*) ∧ manner(go,a) ∧ a=(alone(x))), with (frames, types, roles): $\mathcal{F}$(say), $\mathcal{F}$(decide), $\mathcal{F}$(go), $\mathcal{T}$(Person), $\mathcal{T}$(Place), $\mathcal{R}$(agent), $\mathcal{R}$(time), $\mathcal{R}$(sentence), $\mathcal{R}$(location), $\mathcal{R}$(manner)s

model theory. Unfortunately, an expressive logic of this kind is not appropriate to the current state-of-art applications of web ontologies. On the other hand, KR for the Semantic Web provides compact and tractable languages with a model-theoretic semantics. The main shortcoming is that the strong typing of FRASL must be reconstructed as "meta-level sugar". As an example, (5-13) encode the first part of (4) as a set of OWL2 axioms:

```
(5)    test:sentence_2 frasl:expression:N "Mustafa"[string]
(6)    test:sentence_2 frasl:expression:VP "said"[string@en]
(7)    test:sentence_2 frasl:evokes say_frame:say
(8)    test:say_1 frasl:occurrenceOf say_frame:say
(9)    test:say_1 say_frame:agent test:Mustafa
(10)   test:say_1 say_frame:sentence test:sentence_2
(11)   test:sentence_2 expression:VP "decided"[string^en]
(12)   test:sentence_2 frasl:evokes decide_frame:decide
(13)   test:decide_1 frasl:occurrenceof decide_frame:decide
```

FRASL can be used to describe very different assignment types, e.g. term extraction:

(14) **TermExtraction**{(**extracts**[***agent***:TermExtractor, ***occurrence*****:**"dog" >
…[...(*x*:Dog)], ***corpus***:BNX, ***relevance***:0.7(float)])}

Term extraction, entity resolution and type induction:

(15) **TermExtraction**{(**extracts**[***agent***:NER+ER+SST, ***occurrence*****:**"Immanuel Kant" >
…[dbpedia:Immanuel_Kant(dbo:Person])}

4. FRED as a FRASL application

FRED[3] [14] is a software tool that makes FRASL concrete and applicable to the rapid extraction of frame structures from text. FRED implements some of the constructs described, in particular it reuses several NLP and KR components in order to produce RDF-OWL triples for either predicative or factual structures. For example, FRED is able to produce the RDF graph depicted in Figure 2, extracted from the sentence:

«The statement by China Foreign Ministry on Friday signaled a possible breakthrough in a diplomatic crisis that has threatened American relations with Beijing.»

For comparison, the complete FRASL representation for that sentence would be:

Sentence{"The statement by China Foreign Ministry on Friday signaled a possible breakthrough in a diplomatic crisis that has threatened American relations with Beijing" >
(**signal**[***agent***(*x*:
statement[***agent***:ChinaForeignMinistry(*y*:Organisation)]), ***time***(*t*:past, *t*=Friday)], ***topic***(*y*:
possibility[***event***(e_1:**breakthrough**[***in***(*z*:**diplomaticCrisis**[***event***(e_2:
threaten[***cause***:*z*, ***experiencer***(*w*:
AmericanRelation[***with***:Beijing(Place)])])])])])])])}

Six out of seven frames are detected and represented by FRED (the seventh **possibility** frame requires not yet implemented rules for modality representation).
In addition, FRED provides integration with a named entity recognizer, which resolves one (Beijing) out of two named entities, by linking it to a publicly available multilingual ontology (contextual disambiguation by using inductive classification).

[3] Available at http://wit.istc.cnr.it/stlab-tools/fred

Finally, the conceptual entities extracted can be disambiguated with reference to e.g. WordNet, thus enabling additional conceptual and multilingual interoperability. For example, *statement* can be automatically disambiguated to `wn30:synset-statement-noun-1`, *breakthrough* to `wn30:synset-breakthrough-noun-3`, etc. Disambiguation is also contextual, e.g. with conceptual density or multilingual corpora.

The existence of multilingual ontologies with factual and lexical data (e.g. Wiktionary, DBpedia, WordNet) opens the possibility of rich cross-linguistic queries.

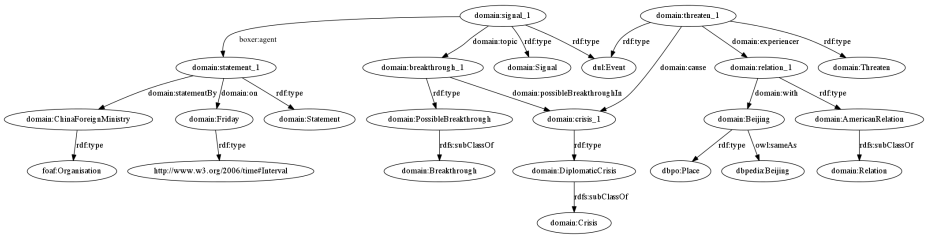

Figure 2: An RDF-OWL graph extracted from the sample sentence by FRED.

References

[1] Gangemi A., Presutti V. A Pattern Science for the Semantic Web, Semantic Web Journal, 1(1), 2010.

[2] Gibson, J. J. The ecological approach to visual perception, Laurence Erlbaum, 1979.

[3] Presutti V., Gangemi A. Identity of Resources and Entities on the Web. International Journal on Semantic Web and Information Systems, 4(2), 2008.

[4] Gruber T. Ontology of Folksonomy: A Mash-up of Apples and Oranges, International Journal on Semantic Web and Information Systems, 3(2), 2007

[5] Abrusci V.M., Romano M. Ontologies, Logic and Interaction: Approaches to Semantic Web ranging from Lexical Semantics to Geometrical Compatibility. Colloque LOCI "Ontologies et Sémantique Lexicale", Roma, 2011.

[6] Beaugrande, R. de. Text, discourse, and process: toward a multidisciplinary science of texts», volume 4 of Advances in discourse processes, ABLEX Pub. Corp., 1980.

[7] Clark P., Thompson J., Porter B. Knowledge Patterns, in Cohn, A.G., et al. (eds.), KR2000, 591-600, Morgan Kaufmann, 2010.

[8] Gangemi A. What's in a Schema?, in Huang, C., Calzolari, N., Gangemi, A., Lenci, A., Oltramari, A., and Prevot, L. (eds.), Ontology and the Lexicon. Cambridge UP, 2010.

[9] Davidson, D. The Logical Form of Action Sentences», in The Logic of Decision and Action (2nd ed.). Pittsburgh: University of Pittsburgh Press, 1968.

[10] Smith, B.C. Levels, layers, and planes. The framework for a theory of knowledge representation semantics. MS Thesis, MIT, 1978.

[11] Fillmore, C., Kay, P., and O'Connor, C. Regularity and Idiomaticity in Grammatical Constructions: The Case of let alone, Language, 64, 501–538 (1988).

[12] Kamp H. A Theory of Truth and Semantic Representation, in J. Groenendijk et al. (eds.). Formal Methods in the Study of Language. Amsterdam: Mathematics Center, 1981

[13] Gangemi A. Back to the Future: Frame Representation and Semantic Technologies. Cahiers de Lexicologie, 2(99), 2011.

[14] Presutti V., Draicchio F., Gangemi A. Knowledge extraction based on Discourse Representation Theory and Linguistic Frames. A. ten Teije and J. Völker (eds.): EKAW2012, LNCS, Springer, 2012.

How the Multilingual Semantic Web can meet the Multilingual Web

A Position Paper

Felix Sasaki
DFKI / W3C Fellow
fsasaki@w3.org

The success of the Web is not based on technology. It is rather based on the availability of tooling to create web content, the fast number of content creators providing content, and finally the users who eagerly "digest" the content and are willing to pay for it, being part of various business models.

Not only the Web in general, but also the Multilingual Web is growing. More and more content is being produced in languages other than English; more and more users want to use their mother tongue on the Web. Unfortunately this growth is not without undesired side effects. "If a language is not on the Web, it doesn't exist" – this phrase[1] expresses the fear of "digital extinction", faced especially by smaller language communities.

To support the Multilingual Web, language technology can play a crucial role: the machine translation of the English Wikipedia articles into Thai is just one example how massive content creation can rely on language technology. The outcome is of course not perfect, and only with human post-editing the result is really useful.

What does all this tell us about the Multilingual Semantic Web (MLSW)? First, like with the Web itself, the availability of standardized technological blocks is a pre-requisite for wide adoption of the MLSW. However, this is not enough. Easy to use tooling to create and to work with RDF based resources is inevitable to lower the barriers for the ordinary content creator. There should be no difference in working with the MLSW compared to editing an HTML web page or setting up a blog.

Second, although the technical infrastructure of the MLSW is given via RDF based building blocks, MLSW resources are rare. Studies[2] reveal that human readable descriptions even in English are hardly available; for other languages or links between languages in the MLSW the situation is even worse.

Third and finally, like for the human readable Web, the application of language technologies can help to create resources for the MLSW, e.g. via the creation of multi-language labels via machine translation. But also like with translation of ordinary Web pages, such approaches need human intervention to assure a certain level of quality.

---

[1] See the presentation from András Kornai at META-FORUM 2012 for details http://www.meta-net.eu/events/meta-forum-2012/report#kornai_presentation

[2] See e.g. the presentation by Jose Emilio Labra Gayo at http://www.multilingualweb.eu/documents/dublin-workshop/dublin-workshop-report#labra

adfa, p. 1, 2011.

The main message of this position statement is that the MLSW has several gaps, which currently hinder the widespread creation and usage of multilingual resources for the Semantic Web. About 2 ½ years ago a similar observation of gaps lead to the creation of a European thematic network, called "The MultilingualWeb"[3]. Via a series of workshops, stakeholders from diverse areas came together and discussed gaps that hinder the development of the Multilingual Web. As one concrete outcome, a EU project was created to develop tooling and standards for a subset of gaps, related to metadata in localization workflows. The W3C MultilingualWeb-LT working group[4] forms the umbrella for this effort. In addition the underlying EU project continues to run the MultilingualWeb workshop series, as a basis for continuous cross-community information exchange and long-term planning.

It seems that for the progress of the MLSW a similar effort is needed. This should not only focus on technology, but on integrating communities. In the remainder of this position statement we will go through the various stakeholder groups identified within the MultilingualWeb workshop series, and will map them to the situation in the MLSW.

Platform Developers provide the technological building blocks that are needed for multilingual content creation and access on the Web. For the Multilingual Web the browser plays a major role. For the MLSW a platform for easy creation of RDF "without seeing the source code" is yet to come. Both the Multilingual Web and the MLSW face challenges in handling of translation workflows. Although more Web content is being translated, the key web technologies HTML and RDF so far have no means to support this process. The forehand mentioned MultilingualWeb-LT working group provides a solution which can be applied to the multilingual Semantic Web as well: upcoming metadata as part of HTML5 based labels in RDF 1.1[5].

The adoption of RDF is hindered by the abstract level of the related standards, lack of outreach, un-harmonized usage of multilingual labeling (see the studies mentioned before), or a lack of testability. A reference implementation of an easy-to-use platform, accompanied by various e.g. educational materials, could boost the adoption of the MLSW. For the Multilingual Web, the W3C has made a long-term effort to raise awareness for multilingual issues via its Internationalization Activity. It is time to work on awareness for the MLSW in this and other fora.

Content Creators more and more need to bring content to different delivery platforms, especially via mobile devices. Since these devices lack computing

[3] See http://www.multilingualweb.eu/ for further information.

[4] See http://www.w3.org/International/multilingualweb/lt/ for further information.

[5] The usage oft the metadata in HTML5 can be seen at http://www.w3.org/TR/its20/#EX-translate-html5 . Since RDF 1.1. encompasses an HTML5 data type, the same approach can be used for translation metadata in RDF labels.

power, many aspects of multilinguality need to be carefully addressed. For the Web in general the creation of applications that work only via the network, e.g. voice analysis and synthesis, has grown. The same holds for the MLSW: device independency can only be achieved if there are stable services which a MLSW "client" can make use of.

The need to create more inter-language links again is valid for both parts of the Web. In the Multilingual Web personalization has become ubiquitous. Search engine providers and other services track user behavior in order to provide the most relevant content in a given situation. The same desire seems to be given for the MLSW: a user e.g. of multilingual RDF resources should not need to have to provide details what parts of the resources (domain specific or general) are relevant; the MLSW "client" should choose the resources based on preferences and tracking of past behavior. Of course such an approach raises privacy issues – and it seems that an initiative like the W3C Tracking Protection working group might then become relevant for the MLSW as well.

The MLSW so far does not address e.g. the requirements of modalities other than text: what role has an image, a video or audio file in the Semantic Web? In the Multilingual Web it is common that such pieces of content are localized to a specific audience – but how about the MLSW? An effort like the English Wikipedia translated into Thai demonstrated the value of combining machine translation with volunteer efforts to create high quality content. For the MLSW, such community approaches are yet to come.

Tooling again seems to be crucial, e.g. to support the easy translation of human readable labels. Explaining the usage of such tooling leads to best practices. For the Web in general, W3C and other organizations recently launched "Web Platform Docs" to provide educational material to a worldwide audience of content creators. Having such material available for the MLSW will be an important step for wider adoption.

Localizers deal with internationalization practice in content creation, the distribution of content to localization companies and the onward distribution to individual translators. Improved efficiency of this process requires technical integration in the resulting workflow.

In this area, the problems of the MLSW are in essence the same as in the Multilingual Web. There is a huge fragmentation of standardization efforts in the localization area. Multiple, sometime overlapping standards are available from different organizations including the W3C, ISO, OASIS, ETSI, or the Unicode consortium. The gap here is often just to understand how the standards interplay.

What does this mean for the MLSW? Truly widespread adoption will mean that Semantic Web resources have to become part of localization workflows and are localized by professional localization companies, by volunteers or a mixture of both. There is no silver bullet to avoid the mistakes being made for

localization of the Multilingual Web. Some advices can be made: not to develop additional standards in this area but rely on existing solutions; integrate localization functionality in a to be developed MLSW platform; and try to match localization workflows, content creators and project needs.

A very promising area in terms of localization tooling seems to be the integration of localization functionality in content creation tools. As mentioned before, the integration of content management and localization is a major task in this area. Bringing MLSW and content creation / localization tooling closer to each other is then just the next logical step.

For **machines**, i.e. applications based on language technology, resources from the MLSW are of (potential or actual) use for cross lingual search, machine translation, multilingual summarization etc. Some language technology applications help to improve the resources of the MLSW, e.g. again machine translation, or data cleansing techniques. The challenges in this area are similar to localization: there are many small solutions, integration has to be done repeatedly, and the re-use of multilingual resources is not straightforward.

Some small integration steps between localization, language technology and the MLSW are being taken. An example is the application of analytics, e.g. named entity annotation, in localization workflows. The dominant format in such workflows is XLIFF (XML Localization Interchange File format). So far there is no standardized way and no tooling available to represent named entities in XLIFF. In the MultilingualWeb-LT working group such tooling is being developed. This will lead to a named entity annotation round tripping workflow from HTML[6] (potentially with an intermediate step via NIF[7]) to XLIFF and back, after translation.

Users normally have no strong voice in the development of multilingual or other technologies. At the MultilingualWeb workshops, it became clear that the worldwide interest in multilingual content is high, but significant organizational and technical challenges need to be tackled for reaching people in less developed economies, especially in linguistically diverse regions such as Asia and Africa. Again, for the creation of content in the MLSW, the same problems apply as well.

A notion that is becoming common in the Multilingual Web is the difference between controlled and uncontrolled environments of content creation and translation. For the MLSW, this seems to be especially crucial for the paradigm of linked open data. Here currently there is practically no difference being made between human language labels created via high quality human translation, or automated results.

---

[6] See an example annotation at http://www.w3.org/TR/its20/#EX-disambiguation-html5-local-1

[7] For details about NIF see http://nlp2rdf.org/nif-1-0

Although the engagement of users is a challenge, it has also promises. A presentation from Facebook at one of the MultilingualWeb workshops revealed that there are 500000 voluntary translators, and that the French instantiation of the site had been translated within 24 hours. A great vision along these lines is a community effort in which fasts amount of content are being created for the MLSW.

Finally, the topic of **policy makers** it is of high importance: many gaps in the Multilingual Web are related to political decisions. Multilingual mandates, participatory democracy or interactive systems for local needs are just a few application scenarios for the MLSW. As a pre-requisite, open multilingual assets are needed, as well as harmonized support across language boundaries. Like with the other areas mentioned in this position statement, such efforts need to be accompanied by education, promotion, coordination, guidelines and business cases related to the MLSW.

Acknowledgements

This position paper was written with input from a paper for the WWW 2012 conference, written by Dave Lewis, David Filip and Felix Sasaki. In addition, a presentation about the outcomes of the MultilingualWeb thematic network given at the TCWorld 2012 conference, given by Christian Lieske and Jan Nelson, provided valuable input. The current MultilingualWeb workshop series receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) Grant Agreement No. 287815.

Cross-lingual Linking on the Multilingual Web of Data (position statement)

Jorge Gracia, Elena Montiel-Ponsoda, and Asunción Gómez-Pérez

Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
{jgracia, emontiel, asun}@fi.upm.es

Abstract. Recently, the Semantic Web has experienced significant advancements in standards and techniques, as well as in the amount of semantic information available online. Even so, mechanisms are still needed to automatically reconcile semantic information when it is expressed in different natural languages, so that access to Web information across language barriers can be improved. That requires developing techniques for discovering and representing cross-lingual links on the Web of Data. In this paper we explore the different dimensions of such a problem and reflect on possible avenues of research on that topic.

Keywords: multilingualism, ontology matching, multilingual linked data, multilingual mappings

1 Motivation

The large and growing amount of semantic data available on the Web, mainly in the form of Linked Data [2], online ontologies, and annotated Web pages, has resulted in the emergence of the so-called Web of Data. This fact has been accompanied by significant advancements in standards and techniques, contributing to the realization of the Semantic Web vision [1]. Some issues, however, need to be solved before a fully realised Semantic Web can be achieved, as for instance, language barriers, amongst others. In this sense, mechanisms are still needed to automatically reconcile semantic data (ontologies and data underlying ontologies) when they are expressed in different natural languages on the Web, in order to enable access to semantic information across language barriers. To this respect, several challenges arise [5], specifically: (i) ontology translation/localization, (ii) cross-lingual ontology linking, (iii) representation of multilingual lexical information, and (iv) cross-lingual access and querying of linked data.

In this paper we focus on the second challenge, namely, the need of establishing, representing, and storing cross-lingual links among semantic information on the Web. In fact, in the multilingual Web of Data that we envision, semantic data with lexical representations in one natural language would be mapped to equivalent or related information in other languages, thus making navigation across multilingual information possible for software agents. In the following we will refer to "cross-lingual ontology linking" in a broad sense, including (semi-) automatic ontology and instance matching methods and techniques applied to the linking of semantic data documented in several natural languages.

The problem of cross-lingual linking is a fundamental one, since more and more legacy data sources available in different natural languages are being transformed into linked data, and have to be linked to be exploited at its full potential. In fact, the establishment of links between or among multilingual data sources would also contribute to the localisation issue, since it would transform monolingual, isolated, data resources into "multilingual resources" just thanks to the links. However, the linking of resources documented in different languages is not so immediate. Several issues that arise in the localization of semantic web resources [4] would be also involved in the liking task, namely, a) conceptualization mismatches due to language and cultural discrepancies; b) conceptualization mismatches due to the perspectives from which the same domain is approached; or even c) different levels of granularity in the conceptualization.

The main purpose of this position paper is to give an insight into the problem of cross-lingual linking on the Web of Data and identify some research topics that will allow us to advance towards a truly multilingual Web of Data. In the rest of the paper (Section 2) we refer to the different knowledge representation levels in which cross-lingual links can be established. Then, we explore the problem and identify possible research lines grouped in three aspects: cross-lingual link discovery, representation, and reuse. Finally, the main conclusions of the paper are summed up in Section 3.

2 Dimensions of the problem and research lines

Cross-lingual links between ontologies and data sources can be established at different knowledge representation levels:

1. Conceptual level: links between ontology entities at the schema level.
2. Instance level: links between data underlying ontologies.
3. Linguistic level: links between lexical representations associated with ontology concepts and/or instances.

The last one is particularly important if certain lexical relations have to be represented across ontologies (e.g., translations or term variations). Each of these levels will require its own link discovery/representation methods and techniques.

In the following we propose some enhancements of available methods and techniques and suggest new avenues of research that could help overcome the problem.

2.1 Cross-lingual Link Discovery

Current ontology matching techniques have to be extended with multilingual capabilities, and novel techniques need to be investigated as well. Cross-lingual links can be discovered by means of some of these techniques:

1. Projecting the lexical content of the mapped ontologies into a common language (either one of the languages of the aligned ontologies or a pivot language) e.g., using machine translation.

2. Comparing the ontology entities directly by means of cross-lingual semantic measures, that is, measures capable of evaluating similarity or relatedness between (ontology) entities documented in different natural languages (e.g., cross-lingual explicit semantic analysis [9]).

Both avenues have to be further explored, compared, and possibly combined. There are a number of early cross-lingual ontology alignment tools that already implement the first technique[1], while the second one remains unexplored yet. Notice that such preliminary systems are intended to discover cross-lingual links at the conceptual level and that cross-lingual alignment systems operating at the instance and linguistic levels are still to come.

An alternate way to discover cross-lingual links is by using the Web of Data as a source of background knowledge. The idea is to infer links from other links already existent among online ontology entities (that are similar to the entities I intend to link). Such an approach was explored in a monolingual context by the Scarlet system [8] and could be extrapolated to a multilingual landscape.

2.2 Cross-lingual Link Representation

In principle, existing constructs of ontology languages can be utilised for representing cross-lingual mappings at the conceptual and instance levels (e.g., owl:sameAs or owl:equivalentClass), whenever the two concepts or instances can be considered cross-lingual equivalents.

Other commonly used vocabularies (e.g. rdfs:subclassOf, skos:narrower or skos:broarder) could also be re-used in case of granularity discrepancies, i.e., when one conceptualization regards a certain concept with a granularity level different from the other conceptualization. In this case, we would suggest an adaptation or enhancement of such relations for a multilingual scenario, so that finer language distinctions are captured.

In the case no equivalence exists (the one language does not conceptualize a certain phenomenon of the world, whereas the other has a concept for it), we could still provide a lexical description for the "inexistent concept" in the target language, provide a link to its closest concept, and signalize it as a specific cross-lingual case. We believe this kind of links should also be accounted for in the Web of Data.

Regarding cross-lingual mappings at the linguistic level, mappings could be established between the natural language descriptions of their concepts. At this level, lexical-semantic relations could be used (hypernym-hyponym, synonym, antonym, translation, etc.). In the simplest case in a cross-lingual scenario, a property labelled "translation" or "cultural equivalent" (for instance) might be established between the lexical realizations of the concepts [7]. Novel ontology lexica representation models [6] have to be explored for this task.

We argue that specific representation models have to be able to define specific relations between natural language descriptions in different languages, what

[1] See for instance the systems that participated in OAEI2011.5 http://oaei.ontologymatching.org/2011.5/multifarm/index.html

we term translation relations or cross-lingual relations. Highly related with this issue is the representation of term variation at a monolingual or multilingual level. A term variant has been defined as "an utterance which is semantically and conceptually related to an original term" [3]. To put it in simple words, we could define them as synonymous terms that refer to the same concept but that highlight a different aspect. We believe that the accounting for and representing term variants would also contribute to the automatic linking of the lexical descriptions associated to concepts (within or across languages).

Further, to facilitate processing and interchange of alignments, specific formats has been proposed in the literature such as the Alignment Format [2] or the EDOAL language [3]. They should be explored and, if needed, extended to accommodate the representation of cross-lingual and multilingual alignments.

2.3 Cross-lingual Link Storage and Reuse

Cross-lingual links can be discovered runtime/offline. However, owing to the growing size and dynamic nature of the Web, it is unrealistic to conceive a Semantic Web in which all possible cross-lingual links are established beforehand. Thus, scalable techniques to dynamically discover cross-lingual links on demand of semantic applications have to be investigated. Although the scalability requirement is not inherent to the multilingual dimension in ontology matching, multilingualism exacerbates the problem due to the introduction of a higher heterogeneity degree and the possible explosion of compared language pairs.

On the other hand, one can imagine some application scenarios (in restricted domains for a restricted number of languages) in which computation and storage of links for later reuse is a viable option. In that case, suitable ways of storing and representing cross-lingual links become crucial. Also links computed runtime could be stored and made available online, thus configuring a sort of pool of cross-lingual links that grows with time. Such online links should follow the Linked Data principles to favour their later access and reuse by other applications.

3 Conclusions

In this paper we have motivated the study of cross-lingual ontology links as one of the fundamental challenges to solve in order to attain the goals of a truly multilingual Web of Data. There are, in particular, three subproblems to treat, namely cross-lingual link discovery, representation, and reuse. We have given an overview of the characteristics of each of them, as well as identified some relevant research topics that have to be further explored to be part of the solution. For instance, representation of cross-lingual links at the linguistic level, as well as the study of cross-lingual semantic measures and cross-lingual ontology alignment techniques. In our view such topics require more atention by the community and

[2] http://alignapi.gforge.inria.fr/format.html
[3] http://alignapi.gforge.inria.fr/edoal.html

will be crucial to enable the multilingual capabilities on the Web of Data.

Acknowledgments. This work is supported by the EU project Monnet (FP7-248458), the Spanish national project BabeLData (TIN2010-17550), and the Spanish Ministry of Economy and Competitiveness within the Juan de la Cierva program.

References

1. T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284(5):34–43, May 2001.
2. C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22, MarMar 2009.
3. B. Daille, B. Habert, C. Jacquemin, and J. Royauté. Empirical observation of term variations and principles for their description. *Terminology*, 3(2):197–258, 1996.
4. M. Espinoza, E. Montiel-Ponsoda, and A. Gmez-Prez. Ontology Localization. In *Proceedings of the 5th International Conference on Knowledge Capture (KCAP09)*, pages 33–40, 2009.
5. J. Gracia, E. M. Ponsoda, P. Cimiano, A. G. Pérez, P. Buitelaar, and J. McCrae. Challenges for the multilingual web of data. *Journal of Web Semantics*, 11:63–71, Mar. 2012.
6. J. McCrae, G. A. de Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gómez-Pérez, J. Gracia, L. Hollink, E. Montiel-Ponsoda, D. Spohr, and T. Wunner. Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*, 46, 2012.
7. E. Montiel-Ponsoda, J. Gracia, G. A. de Cea, and A. Gómez-Pérez. Representing translations on the semantic web. In *Proc. of 2nd Workshop on the Multilingual Semantic Web, at ISWC'11, Bonn, Germany, ISSN 1613-0073*, volume 775, pages 25–37. CEUR-WS, Oct. 2011.
8. M. Sabou, M. d'Aquin, and E. Motta. Exploring the semantic web as background knowledge for ontology matching. *J. Data Semantics*, 11:156–190, 2008.
9. P. Sorg and P. Cimiano. Exploiting wikipedia for cross-lingual and multilingual information retrieval. *Data & Knowledge Engineering*, 74:26–45, Apr. 2012.

The Multilingual Procedural Semantic Web
A Position Paper

Sergei Nirenburg and Marjorie McShane
University of Maryland Baltimore County
sergei@umbc.edu, marge@umbc.edu

The stated goal of the Semantic Web community is to turn the Web into a richly annotated resource, making its content more amenable to applications that involve machine reasoning. The most widely discussed language-oriented aspect of this vision involves the creation and use of an inventory of markup tags that indicate select semantic types. So, the "semantics" of the Semantic Web is not the semantics of full texts or even full sentences, but rather of select elements of text and extra-textual information. Moreover, the annotations are expected to be largely carried out manually, so broad coverage is unlikely, as are consistency and universal public-spiritedness on the part of annotators (cf. Doctorow, no date). Compare this to the ideal semantic web, which would be *automatically* generated from the unadorned web by processors that would carry out lexical disambiguation, referential disambiguation, and the interpretation of textual implicatures, such as the recognition of irony and indirect speech acts. Such full semantic interpretations of web content would serve as optimal input for machine reasoners.

It is common practice in the field of AI to assume the availability of such knowledge structures – in fact, practically all work on machine reasoning over the past decades has used hand-crafted, complete, unambiguous knowledge structures as input. *How* that could be achieved automatically was always considered a separate issue, delegated to the NLP community. The NLP community, however, by and large abandoned the task of deep semantic analysis some 20 years ago, opting to pursue either (a) knowledge lean, "low-hanging fruit" tasks that contribute to the configuration of better NLP applications in the near term but do not contribute to the ultimate goal of automatic text understanding or (b) method-oriented work, in which the methods themselves are of first priority and natural language serves primarily as a source of data sets.[1]

The Semantic Web community has largely followed the spirit of the NLP majority by deeming full semantics to be too complex to be pursed. As such, the semantics of the Semantic Web is effectively constrained to selective annotation of text strings in ways that are considered feasible in the short term. The preferences of the Semantic Web community are reflected in

[1] Space does not permit a full motivation for these generalizations. For that see Nirenburg and McShane, forthcoming as well as the historical references cited therein.

the selection of foci of work: the development of formal standards, metadata tag sets, ontologies to be used as the content of tag sets, and so on. While we appreciate the common preferences of the mainstream NLP and Semantic-Web communities, and while the material below describes an attempt to contribute to the near-term gains they seek, our contributions must be framed within the research paradigm that we deem the most promising for the long-term utility of any NLP, be it for the web or any other corpus: computational deep-semantic processing. We will argue that one near-term results can be achieved within a theory and methodology that seek full understanding of texts, along with associated sophisticated behaviors, by intelligent agents.

Our research program is an outgrowth of the theory of Ontological Semantics, which studies the processes of automatically extracting, representing and manipulating meaning in natural language texts. Analysis by the OntoSem text analyzer pursues all of the desiderata listed in the introductory paragraph, seeking fully specified, unambiguous, ontologically grounded meaning representations that are more amenable to machine reasoning than highly ambiguous natural language texts (Nirenburg and Raskin 2004). Of course, the automatically generated structures are not yet perfect, as that would be well beyond the current state of the art. However, we are making direct progress toward this goal, which suggests that the vision of *fully interpreted content* delivered over the internet should not be neglected. A prototype for this vision was demonstrated in the implemented SemNews application (Java et al. 2007), which took web-delivered news feeds as input and generated semantic interpretations of them represented as RTF structures.

Significantly, Ontological Semantics is a language-independent theory, most of whose knowledge bases (e.g., ontology, fact repository, rule sets for agent decision-making) and reasoning engines are language-independent. In fact, in the intentionally provocatively titled "An NLP lexicon as a largely language independent resource" (McShane et al. 2005), we describe how much of the information found even in the lexicons used to support OntoSem language processing can be directly reused across languages (more on this below). Once the input strings from any language have been interpreted using a battery of processors, the resulting text-meaning representations can be reasoned over by a single set of engines. Language-neutrality offers not only great savings in time for the acquisition of knowledge resources and development of processors, it also offers consistency of processing across languages.

The core point of this statement, which follows basic tenets of configuring intelligent agents within the OntoAgent environment, is as follows. *The only realistic way to enhance the Web with useful semantic annotations is automatically*. Semantic analysis is, by its very nature, procedural: a system – hereafter "agent" – receives some input, analyzes it in context, and generates an interpretation. The component functions of this

process, like all functions, are subject to error; as a result, the agent must be able to evaluate its *confidence* in the function's output based on the overall predictive power of the function as well as the confidence in each input parameter value. Depending upon the calculated confidence in output, the agent can decide whether or not to use the output in a given application. Since many of the actual functions used to generate interpretations are identical (or at least very similar) cross-linguistically, they should be reused to support both efficiency and consistency in the treatment of Web content. Since different functions take different types of parameter values as input – and since some parameter values are quite easy to compute with high confidence while others are much more difficult – it is possible to introduce procedural semantic analyses to web content in a progressive manner, over time.

We will now illustrate how automatic annotations, generated using cross-linguistically applicable functions, could be incorporated into the Semantic Web over time. We will use as sample phenomena so-called indexical expressions, which are strings whose absolute meaning can be understood only with reference to a specific context: e.g., *he, themselves, over there, now, in a few minutes, the preceding paragraph*. The reason why one would want all these indexical expressions fully, locally resolved as annotations to Semantic Web content should be self-explanatory: it is more directly useful to an automatic reasoner to have access to the information "John. W. Lacey III of Kansas City, Kansas died on July 5, 1974 in Washington, D.C. from complications of heart disease" rather than an expression that could be synonymous given the right context: "Yesterday, in that same place, that happened to one of our local boys."

There exists an unfortunate, in our opinion, tradition within the NLP community to treat indexicals in a suboptimal way on at least three fronts. (1) **Unrealistic preconditions.** Most work on automatic pronoun resolution, for example, involves supervised learning (i.e., learning from manually annotated corpora), whose resultant engines require that all future inputs be already annotated, to perfection, in the expected way. (2) **All-or-nothing classifications**. Indexicals are regularly (albeit often tacitly) categorized as "easy" (e.g., *he*) or "too hard" (e.g., pronominal *that*), whereas the actual easy/hard distinction is largely based on the contextual usage of the element. (3) **Language specificity**. Most work on indexicals in NLP and descriptive linguistics is language-specific, but many resolution functions are actually cross-linguistically applicable.[2]

Our proposal is to apply to the Semantic Web the same types of cross-linguistically applicable indexical resolution functions that are already used in the OntoSem environment. The key to successful realization of this proposal

[2] Theoretical work, like that grounded in the tradition of theoretical syntax, typically lacks the needed level of descriptive detail to be of practical utility for NLP.

involves *classifying* usage cases for indexicals with respect to *which parameter values* are required for each decision function and *how* and *with what confidence* those parameter values can be obtained and in each context.

Let us begin by considering some *types, sources* and *confidence levels* of input parameters that might contribute to functions for resolving indexicals found on the Web. **The surface string**: always available, maximally high confidence. **Semantic web annotations**: sometimes available for some types of entities; confidence varies depending on the source, type of tag, etc. **Traditional web annotations**: typically available for html documents; some types of tags (as for formatting) are of high confidence but might be noisy and difficult to automatically interpret. **Automatic "preprocessing" of text**: preprocessing (detecting tokens, proper names, dates, etc.) is a cornerstone of NLP, but web content can be error-prone due to the metadata text, embedded media, etc. **Syntactic analysis**: another mainstream NLP task though even the best current parsers achieve far less than perfect results. **Basic semantic analysis (word sense disambiguation and the determination of dependencies)**: carried out by few NLP systems, OntoSem being among them; analyses tend to be extremely useful in supporting high-level tasks like resolving indexicals, but they are error-prone. Procedural semantic routines to resolve indexicals become more complex, and typically of lower confidence, as they incorporate the latter types of features. But, centrally important for this proposal, the difficulty of each usage case and its associated confidence level can typically be automatically calculated, thus suggesting in which types of applications the automatic results might best be used. Let us consider just a few examples of indexical treatment.

Relative time expressions – such as *today*, *now*, *three weeks from tomorrow* and *in a little while* – can readily be resolved to real times (month, day, year, etc.) if (a) the "anchor time" – i.e., the time of the post (article, etc.) – is known, and (b) the time expression is used outside of direct speech. The former is expected to be recorded in Semantic Web tags, and the latter can typically be determined with high confidence using a preprocessor. (If the expression is within direct speech, then the time of speech must be determined, which requires semantic analysis.) Within OntoSem, the actual functions that can calculate, e.g., *today* vs. *three weeks from tomorrow* are recorded in the "meaning procedures" zones of the respective lexicon entries (McShane et al. 2004). As mentioned earlier, OntoSem lexicons are largely language-independent, meaning that their semantic descriptions and procedural semantic routines can be reused across languages (McShane et al. 2005); so the procedure already available for English *today* can be used to derive the full meaning of Czech *dnes* or Hebrew מהיו – assuming, of course, that preprocessors for these languages are available.

A similar example is the pronoun *I*, which can be resolved with high confidence in one of two cases: (1) if it is used outside of direct speech and

the piece has a single author as indicated by a Semantic Web tag or (2) it is used within an instance of direct speech that contains a preceding instance of *I*. In this latter case, although the real-world referent cannot be confidently distinguished, the coreference relationship between instances of *I* can be. Now contrast *I* with its plural counterpart *we*. *We* is substantially more difficult to interpret since a single author often affirms group membership – explicitly or implicitly – then subsequently speaks on behalf of the group. Alternatively, a piece can be written by more than one person, with *we* in a given context referring either to a subset of the authors or to a larger community to which they all, or a subset of them, belong. The extensive analysis required by people to craft a robust function for resolving *we* underscores why we (yes, we!) should take a cross-linguistic approach to developing procedural semantic functions for the web: it will save the community time and foster consistency of interpretations. Our initial work on the resolution of *we* within OntoSem includes subfunctions for resolving *I* and *we* that involve different kinds of heuristic evidence, some of which we can expect to be available for any language in the short term and other aspects of which require full-blown semantic analyses of the type we are working toward.

Let us conclude by stating that there are many more largely cross-linguistically applicable procedural semantic routines beyond indexicals, for example, the procedure for resolving *very* (as applied to different types of expressions) are (McShane et al. 2004).

References Cited

Doctorow, C. (No date) Metacrap: Putting the torch to seven straw-men of the meta-utopia. Available at http://www.well.com/~doctorow/metacrap.htm

Java, Akshay, Sergei Nirenburg, Marjorie McShane, Timothy Finin, Jesse English, Anupam Joshi. 2007. Using a natural language understanding system to generate Semantic Web content. International Journal on Semantic Web & Information Systems, 3(4), 50-74. October-December 2007.

McShane, Marjorie, Sergei Nirenburg and Stephen Beale. 2005. An NLP lexicon as a largely language independent resource. Machine Translation 19(2): 139-173.

McShane, Marjorie, Stephen Beale and Sergei Nirenburg. 2004. Some meaning procedures of Ontological Semantics. Proceedings of LREC-2004.

Nirenburg, S. & McShane, M. Forthcoming. *Natural Language Processing.* To appear in S.Chipman (ed.) The Oxford Handbook of Cognitive Science.

Nirenburg, S. & Raskin, V. 2004. *Ontological Semantics.* The MIT Press.