

# Preface

The amount of Internet users speaking native languages other than English has seen a substantial growth in recent years. Recent statistics in fact show that the number of non-English Internet users is almost three times the number of English-speaking users. As a consequence, the Web is turning more and more into a truly multilingual platform in which speakers and organizations from different languages and cultural backgrounds collaborate, consuming and producing information at a scale without precedent. Originally conceived by Berners-Lee et al. (2001) as “an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation,” the Semantic Web has seen an impressive growth in recent years in terms of the amount of data published on the Web using the REsource Description Framework (RDF)<sup>1</sup> and OWL<sup>2</sup> data models. The kind of data published on the Semantic Web or Linked Open Data (LOD) cloud is mainly of a factual nature and thus represents a basic body of knowledge that is accessible to mankind as a basis for informed decision-making. The creation of a level playing field in which citizens from all countries have access to the same information and have equal opportunities to contribute to that information is a crucial goal to achieve. Such a level playing field will also reduce information hegemonies and biases, increasing diversity of opinion. However, the semantic vocabularies used to publish factual data in the Semantic Web are mainly in English, which creates a strong bias towards this language and English-based cultures.

As in the traditional Web, language represents an important barrier for information access in the Semantic Web as it is not straightforward to access information produced in a foreign language. A big challenge for the Semantic Web therefore is to develop architectures, frameworks, and systems that can help in overcoming

---

<sup>1</sup><http://www.w3.org/RDF/>.

<sup>2</sup><http://www.w3.org/OWL>.

language and national barriers, facilitating the access to information produced within different cultures and languages. An additional problem is that most of the information on the Web stems from a small set of countries where major languages are spoken. This leads to a situation in which the public discourse is mainly driven and shaped by contributions from those countries where these major languages are spoken. The Semantic Web vision bears an excellent potential to create a level playing field for users with different cultural backgrounds and native languages and originating from different geopolitical environments, the reason being that the information available on the Semantic Web is expressed in a language-independent fashion and thus bears the potential to be accessible to speakers of different languages if the right mediation mechanisms are in place. However, so far the relation between multilingualism and the Semantic Web has not received enough attention in the research community. The goal of this book is therefore to document the state of the art with respect to the above vision of a *Multilingual Semantic Web*, in which semantic information is accessible in multiple and across languages.

The *Multilingual Semantic Web*, as envisioned in this book, would allow for the following functionality:

- Answering information needs in any language with respect to semantically structured data available on the Semantic Web and Linked Open Data cloud
- Verbalizing and accessing semantically structured data, ontologies, or other conceptualizations in different languages
- Harmonization, integration, aggregation, comparison, and repurposing of semantically structured data across languages
- Aligning and reconciling ontologies or other conceptualizations across languages

This book has to some extent been the result of a Dagstuhl Seminar on the “*Multilingual Semantic Web*,” co-organized by Buitelaar et al. in September 2012. Several of the authors of the book chapters were present at this seminar.

The book is divided into three main parts: *Principles*, *Methods*, and *Applications*. The part on Principles discusses formalisms for building the Multilingual Semantic Web. The part on Methods describes algorithms for the construction of the Multilingual Semantic Web. The part on Applications describes the use of the Multilingual Semantic Web in the context of several real-life systems.

## Principles

The chapter by Hirst analyzes the original vision of a Semantic Web by Berners-Lee et al. (2001) and discusses what this vision implies for a Multilingual Semantic Web and the barriers that the nature of language imposes on it. The chapter essentially argues for the impossibility to represent knowledge interlingually by a symbolic language and argues for the exploitation of distributional semantics to represent multilingual content. In particular, the chapter contrasts a writer-oriented and a reader-oriented perspective of the Semantic Web, arguing that so far the

Semantic Web has focused on a writer-perspective and neglected issues related to the perspective of a reader who consumes information on the Semantic Web.

McCrae and Unger describe work at the ontology–lexicon interface and address the issue of how conceptual schemas and RDF datasets can be enriched with linguistic information to express how the elements of the data model can be expressed in different languages. In their work, they build on the *lemon* model and present a domain-specific representation language that builds on patterns to facilitate the creation of *lemon* lexica. This work will thus facilitate the enrichment of the Semantic Web with a lexical layer. They present the creation of a lexicon for DBpedia in English as a use case.

León Araúz and Faber discuss principled issues related to ontology localization. They argue that a lexical layer for the Semantic Web needs to have a suitable formalism for representing and handling cross-lingual variation including syntactic, lexical, conceptual, and semantic features but most importantly also contextual features that model which translation is appropriate in which context. Further, they also present a taxonomy of different types of cross-language equivalence relations.

Pretorius discusses in her chapter the opportunities that the vision of a Multilingual Semantic Web creates for under-resourced languages, in particular for the preservation of indigenous knowledge and thus cultural diversity. In her chapter, she takes a closer look at the challenges that under-resourced languages, in particular South African languages, face. She presents three use cases in which different types of linguistic resources, ranging from multilingual terminologies, indigenous knowledge on astronomy to a parallel corpus based on the South African constitution, are defined and made available as Linked Data.

van Grondelle and Unger present a paradigm for developing scoped human language technology (HLT) applications in the sense that these applications are aligned with a particular application domain and language. They propose a modular architecture for developing HLT applications by decomposing grammars into different modules that can be flexibly composed together in developing a specific application. With this approach, the development of HLT applications is facilitated by a plug-and-play philosophy, and the reuse of components and modules across applications is maximized. A proof-of-concept implementation of this architecture is presented.

Demey and Heath discuss issues related to the verbalization of n-ary relations given that popular Semantic Web formalisms natively support only binary relations. They propose an approach based on reification, which transforms n-ary relations into a set of binary relations. The authors discuss the case of English and Chinese and present a number of typical and representative verbalization patterns for n-ary relations.

## Methods

Vila-Suero et al. are concerned with the challenges in publishing Multilingual Linked Data. They present a methodology for the publication of Multilingual Linked Data that consists of the following steps: (1) specification, (2) modeling, (3) generation, (4) interlinking, and (5) publication. For each of these steps, they discuss aspects, issues, and design decisions, taking into account the multilingual nature of the data.

Alignment of ontologies or conceptualizations originating from different languages is a crucial research topic in the field of the Multilingual Semantic Web. Trojahn et al. discuss the state of the art in cross-lingual ontology matching. On the one hand, they formally define the problem, distinguishing the case of monolingual, multilingual, and cross-lingual ontology matching. On the other, they provide an overview of existing solutions and evaluation datasets and discuss the results of different tools on standard benchmarking datasets.

In a similar vein, Cabrio et al. analyze the synchronization level between language versions of DBpedia. They compare the coverage of the different DBpedia versions with respect to each other, concluding that the versions clearly vary in their completeness, granularity, and coverage, but complement each other. Further, they present an automatic approach to align the properties of different DBpedia language versions and show how these mappings can be exploited in the context of a cross-language question answering system, QAKIS.

Embley et al. present the ML-OntoES system, a semantic search system that supports searching information across languages by mapping them into a language-independent ontology that is shared across languages and into which content in different languages is mapped. A prototype implementation of this paradigm is discussed and shown to deliver satisfactory results.

A crucial aspect of the Multilingual Semantic Web is to enable different stakeholders to engage together and synchronize in the task of developing a joint conceptualization of some domain of common interest. Bosca et al. present an approach along these lines, based on the Moki toolkit, that allows experts to collaborate in creating and translating ontologies across languages. The features that support collaborative ontology management are discussed, focusing on challenging issues and their solution.

An important task within the Multilingual Semantic Web is to move from data models to linguistic representations (generation) and back (interpretation). Gerber and Ngonga Ngomo present a principled approach that is based on BOotstrapping linked data (BOA), a framework that supports the extraction of RDF data from text by inducing a set of lexical patterns. BOA can be used to extract RDF triples from text but also to generate linguistic descriptions from existing triples. A nice feature of BOA is that it follows a language-independent approach and thus can be adapted to different languages straightforwardly. The authors demonstrate the applicability of their approach across languages by training on four different corpora in two different languages (German and English). Further, they show how BOA can

be applied in different applications, e.g., in the task of extracting facts with high accuracy from textual data as well as in the task of validating RDF facts using textual data and in the context of the question answering system Template - based SPARQL Learner (TBSL).

Along similar lines, Damova et al. present an approach that allows one to query semantic knowledge bases in natural language and obtain results from the knowledge base as coherent texts. The solution builds on the Grammatical Framework and implements several transition steps to move from natural language to SPARQL and from a set of RDF triples to coherent natural language text in multiple languages.

Gromann and Declerck address the issue that labels in ontologies are often impoverished by sacrificing linguistic expressivity and completeness for compactness. However, in this way, domain semantics is lost, e.g., through ellipsis. They present a method to expand condensed labels by inferring implicit content from occurrences of ellipsis, which relies on cross-language comparison of labels.

Bond et al. present an approach to develop multilingual lexica linked to a formal ontology. The method is instantiated for WordNet, Global WordNet, and SUMO to create a rich Web of linguistic data linked to axiomatized knowledge.

Tanev and Zavarella present a semiautomatic, weakly supervised approach for lexical acquisition that is language independent and relies on the principle of distributional semantics. It learns semantic classes, modifiers, and event patterns from an unannotated text corpus. The authors discuss the application of this method to reports of natural disasters in Spanish and English.

## Applications

Cross-language and cross-border integration of knowledge is an important topic of research within the Multilingual Semantic Web. An important use case for this is the integration of financial information across countries and legal jurisdictions, in particular business reports that are typically created relative to financial taxonomies used in each country. The eXtensible Business Reporting Language (XBRL) has standardized the generation of and the access to financial statements like balance sheets, but language and XBRL-taxonomy diversity makes financial data integration across national borders and jurisdictions problematic. Integrating financial data in these circumstances requires that different multilingual jurisdictional taxonomies be aligned by finding correspondences between concepts. Thomas et al. present a method to align XBRL taxonomies originating from different countries. The method relies on semantic tagging of accounting concepts, thus narrowing down the possible mappings to a subset of all possible one-to-one mappings.

Thurmair presents an approach to acquire relevant domain knowledge and multilingual terminologies to support ontology-based search across languages. The chapter describes an effort in enhancing an existing system by a natural language query interface in which users can type in a free text query rather than navigate the

ontology to find relevant texts. The acquired multilingual terminologies are used to map a free-text query to the relevant ontology concepts, thus supporting multilingual search. A proof of concept of the ontology-based search approach is provided for the domain of assistive technology.

Murakami et al. present a service-oriented architecture that fosters the easy development of multilingual NLP services and enhances interoperability of language services and facilitates their composition. The chapter describes the architecture of the Language Grid and describes how the service domain model can be used to define service interfaces and service profiles.

## Acknowledgments

This book could not have been written without the support of the EU FP7 program in the context of the projects Monnet (Grant no.: 248458), LIDER (610782), EuroSentiment (296277), and Portdial (296170); the Science Foundation Ireland for the projects Lion2 (SFI/08/CE/I1380) and Insight (SFI/12/RC/2289); and the Deutsche Forschungsgemeinschaft (DFG) via the Excellence Center Cognitive Interaction Technology (CITEC).

We hope that you enjoy the book!

Galway, Ireland  
Bielefeld, Germany  
Spring 2014

Paul Buitelaar  
Philipp Cimiano

## References

- Berners-Lee, T., Hendler, J., & Lassila, O. (2001, May). The semantic web. *Scientific American*, 284(5), 34–43.
- Buitelaar, P., Choi, K. -S., Cimiano, P., & Hovy, E. H. (2012). The multilingual semantic web (Dagstuhl Seminar 12362). *Dagstuhl Reports*, 2(9), 15–94.

Towards the Multilingual Semantic Web

Principles, Methods and Applications

Buitelaar, P.; Cimiano, P. (Eds.)

2014, XV, 333 p. 83 illus., 32 illus. in color., Hardcover

ISBN: 978-3-662-43584-7