Preface

Exploitation of theoretical results in knowledge representation, language standardization by W3C and data publication initiatives such as Linked Open Data have definitively given concreteness to the field of ontology research. In light of these outcomes, ontology development has also found its way, benefiting from years of R&D on dedicated development tools.

However, while basic development and management technologies have reached a wide consensus in both academia and industry, those "more intelligent" aspects focused on how to automate these processes, how to reuse existing resources (from raw text to structured / linguistic resources) to improve existing knowledge, and how to properly interact with different kind of users, are failing to reach industry-standard. Despite interesting and promising results from the area of ontology learning, scientifically proven both on quality and performance of algorithms and on user perspective, there is a daily evidence that "ontologists" are not really exploiting these results, and support from robust and usable tools is quite far from being available.

The next quantum leap in ontology research should thus properly address these high-level aspects: resource reuse (linguistic resources, thesauri etc.), enrichment of contents, networking, support for collaboration between different experts (domain experts, ontologists, engineers, etc.) and knowledge acquisition from text.

This book aims to provide relevant theoretical frameworks and the latest empirical research findings in the ontology development and knowledge acquisition areas. It has been thought and written for researchers willing to find new scientific approaches on knowledge acquisition and management as well as for professionals who want to improve their understanding of these aspects.

The book is organized into three main topics: "Knowledge Acquisition Systems," "Resource Adoption and Reuse to Build Ontologies," and "Semantic Repositories and Relevant Resources Supporting Ontology Development," representing smoothly distinct aspects of Knowledge Acquisition and Evolution. For each of them, we present relevant contributions from researchers and practitioners in the area.

SECTION 1: KNOWLEDGE ACQUISITION SYSTEMS

The first part of this book presents some examples on the current state-of-the-art on systems for automatic knowledge acquisition and ontology development. Different aspects in the realization of such systems, covering methodologies, technological choices, workflow management, and interoperability are being considered under different perspectives and approaches in the first three chapters.

In Chapter 1, "Ontology based Information Extraction under a Bootstrapping Approach," Iosif, Petasis, and Vangelis present a system for acquiring ontological knowledge from multimedia content. The main contributions of their system lies in the combination of machine learning and reasoning approaches, and their bootstrapping framework, where the extraction mechanisms and the evolution of ontology can affect each other in a continuous loop of learn-extract-learn: new acquired data thus feeds the ontology which is evolved over time, and the new entries in the ontology are in turn fed to traditional a IE system, composed of Named Entity Recognizers and Classifiers (NERC) and co-reference processors. This synergic approach is a key feature for improving automatism in learning systems and for fine tuning them to a given domain by adaptively tailoring their behavior on the basis of the same data which represents that domain.

A similar approach to the previous one has been followed by Davide Eynard, Matteo Matteucci, and Fabio Marfia. In Chapter 2, "A modular Framework to Learn Seed Ontologies From Text," they describe their *Extraction* system which aims at producing seed ontologies starting from a corpus of documents relevant to their domain of interest. The ontologies are then used to bootstrap further knowledge acquisition processes by providing core terms and relations.

One characteristic of Extraction is its modularity, as it allows for dynamic pluggability of different and interchangeable methods/strategies for corpus indexing, term selection, hierarchy, and relationship discovery. The system also provides post-analysis viewing utilities, which can support the user even beyond the automatic synthesis performed by the system in a computer aided, though human-centered, process of ontology refinement.

A different perspective on the same task is offered in Chapter 3, "SODA: A Service Oriented Data Acquisition Framework," where Andreea Diosteanu, Armando Stellato, and Andrea Turbati completely focus on the architectural and managerial aspects of semi-automatic knowledge acquisition systems. By taking assessed architectures as a starting step and models for Unstructured Information Management which have reached industry-standard, the authors go beyond these results by proposing an Architecture (and an associated framework) for Computer-Aided Ontology Development (CODA). Former results for data acquisition from unstructured information are thus acknowledged and integrated in the overall architecture oriented to the development and acquisition of ontological data. By relying on these results and in their open interconnectivity, the authors finally present an open solution for rapid development and deployment of services for ontological knowledge acquisition.

SECTION 2: RESOURCE ADOPTION AND REUSE TO BUILD ONTOLOGIES AND SEMANTIC REPOSITORIES

The heterogeneity of available sources that can be processed to feed semantic repositories is not only limited to unstructured information access: the lack of explicit semantics in many data repositories (from databases to xml storages) makes them as unknown and unintelligible as any unstructured content. However, common sense and universally adopted patterns in modeling knowledge in existing data structures may lead to "guesses" which can be analyzed, tested, and verified.

In Chapter 4, "Mining XML Schemas to Extract Conceptual Knowledge," Ivan Bedini, Benjamin Nguyen, Christopher Matheus, Peter F. Patel-Schneider, and Aidan Boran present the result of their detailed analysis and classification work on patterns for automatic transformation of XML Schemas into RDF and OWL. The many variations in XML schema do not automatically (and univocally) imply given

semantic patterns in ontology modeling, and an attempt to find such mappings between XML structures and RDF/OWL ends in the discovery of a n-to-n search space. Despite the discouraging premises, statistical expectances, cross-checking with available resources, and language analysis lay the constraints which may help to disambiguate this process, which seems to require more craft than analytical skills.

While XML may be seen as a less noble form of ontological data, where the semantics of data are implicit in the structure of the tags which has been thought by the author of the XMLSchema, there are also other resources where the semantics are more explicit. But it is their content which does not directly offer the same perspective on the information that is intended to be represented in a domain ontology. In those cases, semiautomatic processes for knowledge acquisition have to be highly informed about these different perspectives, and thus be able to extract, project, and eventually modify the information coming from the source to produce suitable ontologies/data.

In Chapter 5, "LMF Dictionary-Based Approach for Domain Ontology Generation," Feten Baccar Ben Amar, Bilel Gargouri, and Abdelmajid Ben Hamadou propose an approach for generating domain ontologies from Machine Readable Dictionaries and Lexicons written through the Lexical Markup Framework (LMF). The domain of the two worlds is different: ontologies are made of concepts, relations, and objects of the world, LMF dictionaries contain the words used to the describe these concepts, objects, and relations. Yet, both these domains may share a significant overlap on a common domain of discourse, and, after all, ontologies need natural language to be really shared by humans upon any sort of real-world interpretation dictated by common sense and common knowledge. With this principle in mind, the authors try to guide the development of ontologies through sets of evidence, common patterns, etc. provided by LMF lexicons and by defining a processing chain starting from the identification of the domain (and thus the cut to apply to the LMF dictionary), the selection of concepts, and the progressive enrichment of the ontology.

In Chapter 6, "OntoWiktionary: Constructing an Ontology from the Collaborative Online Dictionary Wiktionary," Christian Meyer and Iryna Gurevych describe their approach for constructing ontologies based on the extraction of terms from Wiktionary, a collaborative online dictionary encoding information about words, word senses, and relations between them. The collaborative nature of this resource, which closely recalls the Wikipedia approach (and comes in fact from the same Wikimedia Foundation that created Wikipedia), and the finer semantic organization of its entries with respect to its encyclopedic sibling, provides a fertile ground for building domain ontologies. In the same chapter, the authors report on the development of OntoWiktionary, an ontology which has been entirely derived by harvesting data from Wiktionary and by "ontologizing" this information.

The last frontier of "ontology development by reuse of existing resources" lies probably in the development of ontologies based on aggregation of other ontologies! or at least, that's what Mariana Damova, Atanas Kiryakov, Maurice Grinberg, Michael K. Bergman, Frédérick Giassonm, and Kiril Simov thought when they initiated their work (presented in Chapter 7, "Creation and Integration of Reference Ontologies for Efficient LOD Management") on the Reference Knowledge Stack (RKS). The RKS is thought of as a reference point for access to LOD data, where general upper ontologies with progressively more detailed levels of conceptualization are interconnected and made available to users, and where reasoning is made possible by means of reasonable views, a form of local pre-processing, matching, cleaning, and reasoning of given sections of the LOD (where "local" may still mean billions of triples). The chapter presents the methods (manual and semi-automatic) used in the creation of the RKS and provides examples illustrating advantages in the use of RKS for managing highly heterogeneous data and its usefulness in real life knowledge intense applications.

SECTION 3: RELEVANT RESOURCES SUPPORTING ONTOLOGY DEVELOPMENT

If resource reuse is an important aspect of ontology development, then there can be specialized resources addressing different aspects of the representation of knowledge that can be thought of as effectively reusable and interlinkable modules for domain ontologies. Proper representation of the linguistic aspects of information is one of these aspects which is often underestimated when one has to focus on the proper conceptualization for a given domain. So, why separate the two aspects: conceptualization and linguistic description, and provide rich and linguistically motivated resources, which can then be properly connected to entries in domain ontologies?

In Chapter 8, "Aggregation and Maintenance of Multilingual Linked Data," Ernesto William De Luca explores issues related to the aggregation and maintenance of Multilingual Linked Data, and how a proper linguistic characterization of ontologies may greatly support search and personalization in user-tailored systems.

Wikipedia is a semi-structured resource, as it features lot of free-text which is somewhat organized (there is a plethora of Wikipedia templates on how to write articles for specific themes, which bring explicit semantics at least to the structure of the article, and also help to disambiguate the terms inside it) and sometimes flanked by explicit semantic tags. As a consequence, elicitation of semantic content from its free-text is expected to be facilitated by the surrounding semantic context and the extracted information should be of high quality.

In Chapter 9, **"**Mining Multiword Terms from Wikipedia," Silvana Hartmann, György Szarvas, and Iryna Gurevych present their research work on the first necessary step for every automatic ontology development process: the extraction of terminology. In particular, they focus on the extraction of multiword terms, which are poorly represented in standard lexical resources, but which typically express explicit concepts on their own. The presented method thus benefits from the underlying semantic structure of Wikipedia and from the huge quantity of information which it provides as well.

The last chapter of this book, "Exploiting Transitivity in Probabilistic Models for Ontology Learning," by Francesca Fallucchi and Fabio Massimo Zanzotto, focuses on methods and techniques for incremental ontology learning: probabilistic methods to learn information for a specific domain by exploiting seed ontologies in more generic domains and solutions (supported by their development and integration inside a graphic ontology editing and knowledge acquisition tool) for putting "human feedback in the middle" of a statistical learning loop.

This book could not have been realized without the constant help and support of its Editorial Advisory Board, composed of an outstanding group of people who have greatly contributed to both the academia and industry in this field of research. Our thankful wishes thus go to Aldo Gangemi (Consiglio Nazionale delle Ricerche – CNR, Italy), Francesco Guerra (University of Modena and Reggio Emilia, Italy), Dickson Lukose (MIMOS, Malaysia), Diana Maynard (Sheffield University, UK), John McCrae (University of Bielefeld, Germany), Frederique Segond (Xerox Research Centere Europe, France), Michael Uschold (Semantic Arts, USA) and René Witte (Concordia University, Canada) for supporting this endeavor with their scientific reviewing and in many other ways. We would also like to send our kudos to the other people who kindly volunteered for supporting the review work: Éric Charton, Aaron Kaplan, Nikolaos Lagos, Marie-Jean Meurs, Alexandre Riazanov, Claude Roux, and Andrea Turbati.

We are grateful to the IGI Global Team for their support and assistance along the long path which leads to the publication of a new book.

Finally, a big "thank you" to the authors, who have contributed with their work and dedication, bringing interesting and novel approaches and ideas in this field of study. We thank you for your contribution to research.

Maria Teresa Pazienza
University of Roma Tor Vergata, Italy

Armando Stellato
University of Roma Tor Vergata, Italy

October, 2011