Editorial Advisory Board

Christopher Brewster, Aston University, UK
Chunyu Kit, City University of Hong Kong, China
Philipp Cimiano, University of Bielefeld, Germany
Sophia Ananiadou, University of Manchester, UK
Tharam Dillon, Curtin University of Technology, Australia
Venkata Subramaniam, IBM India Research, India
# Table of Contents

**Foreword** ............................................................................................................................................. xii

**Preface** ................................................................................................................................................ xiv

**Acknowledgment** ............................................................................................................................. xvii

## Section 1
**Techniques for Ontology Learning and Knowledge Discovery**

**Chapter 1**
Evidence Sources, Methods and Use Cases for Learning Lightweight Domain Ontologies........... 1

*Albert Weichselbraun, Vienna University of Economics and Business, Austria
Gerhard Wohlgenannt, Vienna University of Economics and Business, Austria
Arno Scharl, MODUL University, Austria*

**Chapter 2**
An Overview of Shallow and Deep Natural Language Processing for Ontology Learning.......... 16

*Amal Zouaq, Simon Fraser University - Athabasca University, Canada*

**Chapter 3**
Topic Extraction for Ontology Learning .............................................................................................. 38

*Marian-Andrei Rizoiu, University Lumière Lyon 2, France
Julien Velcin, University Lumière Lyon 2, France*

**Chapter 4**
A Cognitive-based Approach to Identify Topics in Text Using the Web as a Knowledge Source ...... 62

*Louis Massey, Royal Military College of Canada, Canada
Wilson Wong, University of Western Australia, Australia*

**Chapter 5**
Named Entity Recognition for Ontology Population Using Background Knowledge from Wikipedia .................................................................................................................. 80
Chapter 6
User-Centered Maintenance of Concept Hierarchies

Kai Eckert, University of Mannheim, Germany
Robert Meusel, University of Mannheim, Germany
Heiner Stuckenschmidt, University of Mannheim, Germany

Chapter 7
Learning SKOS Relations for Terminological Ontologies from Text

Wei Wang, University of Nottingham Malaysia Campus, Malaysia
Payam M. Barnaghi, University of Surrey, United Kingdom
Andrzej Bargiela, University of Nottingham Jubilee Campus, United Kingdom

Section 2
Applications of Ontologies and Knowledge Bases

Chapter 8
Incorporating Correlations among Gene Ontology Terms into Predicting Protein Functions

Pingzhao Hu, York University - University of Toronto, Canada
Hui Jiang, York University - University of Toronto, Canada
Andrew Emili, University of Toronto, Canada

Chapter 9
GO-Based Term Semantic Similarity

Marco A. Alvarez, Utah State University, United States
Xiaojun Qi, Utah State University, United States
Changhui Yan, North Dakota State University, United States

Chapter 10
Ontology Learning and The Humanities

Toby Burrows, University of Western Australia, Australia

Chapter 11
Ontology-Based Knowledge Capture & Sharing in Enterprise Organisations

Aba-Sah Dadzie, University of Sheffield, United Kingdom
Victoria Uren, University of Sheffield, United Kingdom
Fabio Ciravegna, University of Sheffield, United Kingdom
Section 3
Emerging Trends in Ontology Learning and Knowledge Discovery

Chapter 12
Automated Learning of Social Ontologies ................................................................. 229

Konstantinos Kotis, University of the Aegean, Greece
Andreas Papasalouros, University of the Aegean, Greece

Chapter 13
Mining Parallel Knowledge From Comparable Patents .............................................. 249

Bin Lu, City University of Hong Kong, Hong Kong
Benjamin Tsou, City University of Hong Kong - Hong Kong Institute of Education, Hong Kong
Tao Jiang, ChiLin Star Corporation, China
Jingbo Zhu, Northeastern University, China
Oiyee Kwong, City University of Hong Kong, Hong Kong

Chapter 14
Cross-language Ontology Learning ............................................................................. 274

Hans Hjelm, alaTest.com, Sweeden
Martin Volk, University of Zurich, Switzerland

Compilation of References .......................................................................................... 301

About the Contributors ............................................................................................... 329

Index ............................................................................................................................. 337
Detailed Table of Contents

Foreword .................................................................................................................................................... xii

Preface ........................................................................................................................................................... xiv

Acknowledgment ........................................................................................................................................... xvii

Section 1
Techniques for Ontology Learning and Knowledge Discovery

Chapter 1
Evidence Sources, Methods and Use Cases for Learning Lightweight Domain Ontologies............... 1
Albert Weichselbraun, Vienna University of Economics and Business, Austria
Gerhard Wohlgenannt, Vienna University of Economics and Business, Austria
Arno Scharl, MODUL University, Austria

Section 1 starts off with Chapter 1 presenting a condensed view of the current work in ontology learning, with a particular focus on a variety of emerging knowledge sources, most of which are Web-based, for discovering domain concepts and labelling the relations between them. As a proof of concept, the authors put forward a generic framework for learning lightweight ontologies using unstructured data sources on the Web such as FreeBase, Flickr, Twitter, and Technorati. Three use cases in tourism, waste management, and climate change, respectively, are presented to demonstrate the applicability of the presented framework to real-world problems as well as for discussing possible pitfalls in ontology learning.

Chapter 2
An Overview of Shallow and Deep Natural Language Processing for Ontology Learning............. 16
Amal Zouaq, Simon Fraser University - Athabasca University, Canada

This chapter complements the review in Chapter 1 by focusing on natural language processing techniques for ontology learning, with a particular emphasis on deep semantic analysis methods. The authors discuss both shallow as well as deep techniques for natural language processing, and describe their applicability to each step of learning ontologies from text documents, especially those from the
Web. In particular, the authors discuss the use of lexico-syntactic patterns based on dependency grammars coupled with Web resources for learning ontologies, and explain why this move can be considered as a step towards deeper semantic analysis in ontology learning.

Chapter 3
Topic Extraction for Ontology Learning ................................................................. 38

Marian-Andrei Rizoiu, University Lumière Lyon 2, France
Julien Velcin, University Lumière Lyon 2, France

This chapter initiates the transition from the lexical layer in Chapter 2 to the concept layer of ontologies by looking at the application of topic extraction techniques for ontology learning. The authors provide a thorough review of clustering and term extraction techniques, and describe how these techniques are assembled together into an unsupervised learning system for extracting meaningful topics. The chapter also includes some results from an initial experiment with system using the Reuters-21578 corpus. While the extracted topics are not full concepts as yet, they nevertheless provide an excellent starting point for constructing concepts in ontology learning.

Chapter 4
A Cognitive-based Approach to Identify Topics in Text Using the Web as a Knowledge Source .... 62

Louis Massey, Royal Military College of Canada, Canada
Wilson Wong, University of Western Australia, Australia

This chapter presents a revolutionary approach for identifying topics from text that is fundamentally different from the existing techniques reviewed in Chapter 3. The authors discuss how the approach, which is inspired by human cognition, allows ‘meaning’ to emerge naturally from the activation and decay of unstructured text information retrieved from the Web. Using the unstructured texts in Web pages as a source of knowledge alleviates the laborious handcrafting of formal knowledge bases and ontologies which are required by many existing techniques. The authors discuss the results from some initial experiments comparing the use of WordNet versus Web pages from Yahoo! search on the Reuters-21578 corpus to illustrate the power of this new approach.

Chapter 5
Named Entity Recognition for Ontology Population Using Background Knowledge from Wikipedia .................................................. 80

Ziqi Zhang, University of Sheffield, UK
Fabio Ciravegna, University of Sheffield, UK

This chapter offers an alternative, classification-based view of concept discovery using named-entity recognition techniques as opposed to the notion of topics in Chapter 3 and 4. The authors provide a thorough review of named-entity recognition techniques with an emphasis on those using the Web for background knowledge. The authors then propose a novel method that automatically creates domain-specific background knowledge by exploring Wikipedia for classifying terms into predefined ontological classes. The authors also demonstrate the potential use of this method for ontology population.
Chapter 6
User-Centered Maintenance of Concept Hierarchies

Kai Eckert, University of Mannheim, Germany
Robert Meusel, University of Mannheim, Germany
Heiner Stuckenschmidt, University of Mannheim, Germany

This chapter extends the discussion on discovering concepts from the perspective of topics and named entities in Chapter 3, 4, and 5 to ascertaining the taxonomic, also known as hierarchical, relations between concepts. The authors describe several approaches for placing new concepts in existing hierarchies, creating new hierarchies by means of crowdsourcing, and visualising hierarchies, all of which are essential to the semi-automatic creation and maintenance of concept hierarchies. The authors also discuss how unstructured text provided by Web search engines, existing structured resources such as Medical Subject Headings (MeSH) and WordNet, and collaborative editing come together as important resources for the above three tasks.

Chapter 7
Learning SKOS Relations for Terminological Ontologies from Text

Wei Wang, University of Nottingham Malaysia Campus, Malaysia
Payam M. Barnaghi, University of Surrey, United Kingdom
Andrzej Bargiela, University of Nottingham Jubilee Campus, United Kingdom

This chapter extends the discussion in Chapter 6 on taxonomic relations to include relations of type “related” between concepts. The authors present a novel approach based on probabilistic topic models using Latent Dirichlet Allocation to automatically learn both the “broader” as well as the “related” type of relations from unstructured text corpus. The authors generated around 7,000 ontological statements expressed in terms of the two types of relations from abstracts of publications in the Semantic Web research area.

Section 2
Applications of Ontologies and Knowledge Bases

Chapter 8
Incorporating Correlations among Gene Ontology Terms into Predicting Protein Functions

Pingzhao Hu, York University - University of Toronto, Canada
Hui Jiang, York University - University of Toronto, Canada
Andrew Emili, University of Toronto, Canada

This chapter begins Section 2 by looking at a popular application of ontologies since the post-genomic era, which is the assignment of functions to uncharacterized proteins. Since proteins mostly interact with other bio-molecular units to execute their functions, the functions of unknown proteins may be discovered by studying their associations with proteins having known functions. This chapter compares neighbourhood-based and global techniques for the automated prediction of protein functions based on the analysis of the patterns of functional associations in interaction networks such as the Gene Ontology.
The authors briefly discuss how the taxonomic structure in GO, or in other words, the interrelations between function terms, is used to compute semantic similarity to enhance predictions using these techniques.

**Chapter 9**
GO-Based Term Semantic Similarity ................................................................. 176

Marco A. Alvarez, Utah State University, United States
Xiaojun Qi, Utah State University, United States
Changhui Yan, North Dakota State University, United States

This chapter proceeds with the discussion raised in Chapter 8 on the use of the Gene Ontology (GO) to compute semantic similarity. The authors first provide a brief review of the current advances and challenges in the development of techniques for calculating semantic similarity between GO terms. The authors then introduce a new technique that exploits the different properties of GO to calculate semantic similarities between pairs of GO terms. In this chapter, the authors combine information about the shortest path between terms, the nearest common ancestor for a pair of terms, and the extent of overlap in the definitions of terms for computing similarity.

**Chapter 10**
Ontology Learning and The Humanities ............................................................ 188

Toby Burrows, University of Western Australia, Australia

This chapter shifts the discussion from the biomedical domain in Chapter 8 and 9 to the use of ontologies in projects and services for the humanities. The author reviews several examples including the EU VICODI project which uses ontologies to aid the searching a European history portal, the CIDOC conceptual model that can be found in a range of museum documentation projects such as CLAROS, which combines large database records for Greek and Roman art objects, and other smaller scale applications of more specific ontologies. The author explains that the slow progress in applying ontology learning techniques in the humanities can be attributed to the complexity of the linguistic and conceptual environment of the domain. The author discusses four recent projects on the learning of ontologies from major collections of classical literature, 18th-century French writings, 13th-century Latin documents and 18th-century accounts of trials.

**Chapter 11**
Ontology-Based Knowledge Capture & Sharing in Enterprise Organisations .................. 202

Aba-Sah Dadzie, University of Sheffield, United Kingdom
Victoria Uren, University of Sheffield, United Kingdom
Fabio Ciravegna, University of Sheffield, United Kingdom

This chapter concludes Section 2 by focusing on the application of ontologies to improve knowledge management in organisations. The authors present a knowledge framework that integrates a number of tools to provide alternative perspectives on data to suit different users and tasks across multiple communities of practice. This framework uses domain ontologies to guide users in data exploration and analysis and support the contextualisation and codifying of information into knowledge. The authors
evaluate the framework with end users in the aerospace engineering domain, using a simulation of an information retrieval case. The participants found that the ontology-guided hypothesis exploration and investigation aided the contextualisation of information, leading to an increase in the confidence with which they came to conclusions about the simulated issues.

Section 3
Emerging Trends in Ontology Learning and Knowledge Discovery

Chapter 12
Automated Learning of Social Ontologies ................................................................. 229
Konstantinos Kotis, University of the Aegean, Greece
Andreas Papasalouros, University of the Aegean, Greece

Section 3 starts off with Chapter 12 that focuses on an emerging data source for learning ontologies. This chapter begins with a discussion on the requirements for the automated learning of ontologies from social data on the Web such as blogs, wikis, and folksonomies. The authors then present two techniques for automatically learning ontologies of social concepts and relations from query logs, and Web 2.0 question/answer applications such as Yahoo! Answer. The authors evaluate the ontology learning technique from query logs using Yahoo! and Google query datasets. The authors also discuss the importance of modelling trust for specifying the degree of confidence that agents, both software and human, may have on the conceptualisations derived from social content.

Chapter 13
Mining Parallel Knowledge From Comparable Patents ............................................. 249
Bin Lu, City University of Hong Kong, Hong Kong
Benjamin Tsou, City University of Hong Kong - Hong Kong Institute of Education, Hong Kong
Tao Jiang, ChiLin Star Corporation, China
Jingbo Zhu, Northeastern University, China
Oiyee Kwong, City University of Hong Kong, Hong Kong

This chapter turns the focus from social data in Chapter 12 to another emerging type of data for ontology learning and knowledge discovery. This chapter discusses the potentials and challenges of using linguistically diverse Web data to address the problem of mining the same knowledge across different languages. The authors focus on the mining of parallel sentences and parallel technical terms from comparable Chinese-English patent texts which contain both equivalent sentences as well as much noise. The authors touch on the potential use of the extracted parallel sentences and technical terms for further acquisition of terms and relations, translation of monolingual ontologies, as well as other cross-lingual information access applications.
Chapter 14
Cross-language Ontology Learning ................................................................. 274

Hans Hjelm, alaTest.com, Sweden
Martin Volk, University of Zurich, Switzerland

This chapter generalises the discussion in Chapter 13, and brings the book to an end by reflecting on the fact that a formal ontology does not contain lexical knowledge and hence, by nature is language independent. This chapter focuses on ways to automatically build ontologies by exploiting cross-language information from parallel corpora. In particular, the authors present a framework that provides a setting in which cross-language data can be integrated and quantified for cross-language ontology learning. The authors employ resources such as the JRC-ACQUIS Multilingual Parallel Corpus and the Eurovoc multilingual thesaurus for their experiments. The authors conclude that the combining of information from different languages can indeed improve the results of ontology learning.

Compilation of References .................................................................................. 301

About the Contributors ........................................................................................ 329

Index ...................................................................................................................... 337
Foreword


In half a generation, we have moved from a world in which it was hard to discover information about any given subject into one where we feel surrounded, almost imprisoned, by more than we could possibly hope to digest. But, paradoxically, it is harder than ever to keep ourselves informed. How can anyone read and process the Web, which is updated and augmented every second?

Herein lies the key: whereas in the old world virtually all information was recorded on paper, now everything is electronic. Recent decades have seen computational linguists join forces with information professionals and computer scientists to develop productive ways of digesting vast quantities of electronic text, whether automatically or under human oversight. New paradigms of language processing have sprung from the ready availability of corpora whose size was unimaginable 20 years ago, giving birth to fields such as text mining and information extraction.

The information is readily available, and we have ways of analyzing it linguistically. But to find out what it means we need knowledge. Today’s bottleneck is in handcrafting structured knowledge sources—dictionaries, taxonomies, knowledge bases, and annotated corpora. Tomorrow’s machines will unravel knowledge from information automatically. And to do so, they will employ one of philosophy’s most fundamental concepts: ontology.

Ontology is the study of the nature of being. It concerns what entities exist and how they can be referred to, grouped together, and categorized according to their similarities and differences. The ontologies used in information science are formal representations of concepts and their relationships with one another. An ontology provides a shared vocabulary that can be used to model a domain and talk about it. The need to relate different pieces of information boils down, in essence, to the deep problem of learning and relating different ontologies. Ontologies have moved from an obscure corner of metaphysics to occupy center stage in the world of information processing.

The time is ripe for this book. Techniques of ontology learning and knowledge discovery are beginning to converge. Prototypes are becoming stronger. Industry practitioners are beginning to realize the need for ontology learning. Wilson, Wei, and Mohammed bring together recent work in the construction and application of ontologies and knowledge bases. They introduce a wide range of techniques that utilize unstructured and semi-structured Web data for learning and discovery.
Section I covers existing and emerging techniques for extracting terms, concepts, and relations to construct ontologies and knowledge bases. It provides a background in natural language processing that moves up from the lexical to the concept layer. Knowledge sources include the Web, Wikipedia, and crowd-sourced repositories. One chapter introduces a new topic extraction technique for concept discovery; another promotes the use of existing deep semantic analysis methods in ontology learning. Section II examines how ontologies and knowledge bases are being applied across different domains: biomedicine, genetics, enterprise knowledge management, and the humanities. Section III focuses on emerging trends: learning ontologies from social network data and improving knowledge discovery using linguistically diverse Web data.

The interdisciplinary nature of ontology learning and knowledge discovery is reflected in this book. It will appeal to advanced undergraduates, postgraduate students, academic researchers and practitioners. I hope that it will lead to a world in which we can all live more effectively, a world in which the ready availability of information is balanced by our enhanced ability to process it.

Ian H. Witten
September 2010
Preface

It has become sort of a cliché nowadays to mention how rapidly textual information is growing and how the World Wide Web has assisted in this growth. This, however, does not shadow the fact that such explosive growth will only intensify for years to come, and more new challenges and opportunities will arise. Advances in fundamental areas such as information retrieval, machine learning, data mining, natural language processing, and knowledge representation and reasoning have provided us with some relief by uncovering and representing facts and patterns in text to ease the management, retrieval, and interpretation process. *Information retrieval*, for instance, provides various algorithms to analyse associations between components of a text using vectors, matrices, and probabilistic theorems. *Machine learning* and *data mining*, on the other hand, offer the ability to learn rules and patterns out of massive datasets in a supervised or unsupervised manner based on extensive statistical analysis. *Natural language processing* provides the tools for analysing natural language text on various language levels (e.g. morphology, syntax, semantics) to uncover manifestations of concepts and relations through linguistic cues. *Knowledge representation and reasoning* enable the extracted knowledge to be formally specified and represented such that new knowledge can be deduced.

The realization that a more systematic way of consolidating the discovered facts and patterns into an organised, higher level construct to enhance everyday applications (e.g. Web search) and enable intelligent systems (e.g. Semantic Web) eventually gave rise to ontology learning and knowledge discovery. *Ontologies* are effectively formal and explicit specifications, in the form of concepts and relations, of shared conceptualisations, while *knowledge bases* can be obtained by populating the ontologies with instances. Occasionally, ontologies contain axioms for validation and constraint definition. As an analogy, consider an ontology as a cupcake mould and knowledge bases as the actual cupcakes of assorted colours, tastes, and so on. *Ontology learning* from text is then essentially the process of deriving the high-level concepts and relations from textual information. Considering this perspective, *knowledge discovery* can refer to two things, the first denotation being the uncovering of relevant instances from data to populate the ontologies (also known as *ontology population*), and the second, more general sense being the searching of data for useful patterns. In this book, knowledge discovery can mean either one of the two.

Being a young and exciting field, ontology learning has witnessed a relatively fast progress due to its adoption of established techniques from the related areas discussed above. Aside from the inherent challenges of processing natural language, one of the remaining obstacles preventing the large-scale deployment of ontology learning systems is the bottleneck in handcrafting structured knowledge sources (e.g. dictionaries, taxonomies, knowledge bases) and training data (e.g. annotated text corpora). It is gradually becoming apparent that in order to minimize human efforts in the learning process, and to
improve the scalability and robustness of the system, static and expert crafted resources may no longer be adequate. An increasing amount of research effort is being directed towards harnessing collective intelligence on the Web as an attempt to address this major bottleneck. At the same time, as with many fields before ontology learning, the process of maturing has triggered an increased awareness of the difficulties in automatically discovering all components of an ontology, i.e. terms, concepts, relations, and especially axioms. This gives rise to the question of whether the ultimate goal of achieving full-fledged formal ontologies automatically can be achieved. While some individuals dwell on the question, many others have moved on with a more pragmatic goal, which is to focus on learning lightweight ontologies first, and extend them later if possible. With high hopes and achievable aims, we are now witnessing a growing interest in ontologies across different domains that require interoperability of semantics and a touch of intelligence in their applications.

This book brings together some of the latest work on three popular research directions in ontology learning and knowledge discovery today, namely, (1) the use of Web data to address the knowledge and training data preparation bottleneck, (2) the focus on lightweight ontologies, and (3) the application of ontologies in different domains and across different languages. Section I of the book contains chapters covering the use of a wide range of existing, adapted and emerging techniques for extracting terms, concepts and relations to construct ontologies and knowledge bases. For instance, in addition to traditional clustering techniques reported in Chapter III, a new topic extraction technique is being devised as in Chapter IV to offer alternative ways for discovering concepts. Chapter II, on the other hand, promotes the new application of existing deep semantic analysis methods for ontology learning in general. The use of semi-structured Web data such as Wikipedia for named entity recognition, and the question of how can this be applicable to ontology learning are also investigated in Chapter V. The focus of Chapter I is on the construction of practical, lightweight ontologies for three domains. As for Chapter VI and VII, the authors mainly investigate the use of a combination of data sources, both local and from the Web, to discover hierarchical and non-taxonomic relations. In Section II, the authors look at how ontologies and knowledge bases are currently being applied across different domains. Some of the domains covered by the chapters in this section include biomedical (Chapter VIII and IX), humanities (Chapter X) and enterprise knowledge management (Chapter XI). This book ends with Section III that covers chapters on the use of social data (Chapter XII) and parallel texts (Chapter XIII and XIV), which may or may not be from the Web for learning social ontologies, incorporating trust into ontologies, and improving the process learning ontologies.

This volume is both a valuable standalone as well as a great complement to the existing books on ontology learning that have been published since the turn of the millennium. Some of the previous books focus mainly on techniques and evaluations, while others look at more abstract concerns such as ontology languages, standards, and engineering environments. While the background discussions on the techniques and evaluations are indispensable, the focal point of this book remains on emerging research directions involving the use of Web data for ontology learning, the learning of lightweight as well as cross-language ontologies, and the involvement of ontologies in real-world applications. We are certain that the content of this book will be of interest to a wide ranging audience. From a teaching viewpoint, the book is intended for undergraduate students at the final year level, or postgraduate students who wish to learn about the basic techniques for ontology learning. From a researcher’s and practitioner’s point of view, this volume will be an excellent addition outlining the most recent progress to complement basic references in ontology learning. A basic familiarity with natural language processing, probability
and statistics, and some fundamental Web technologies such as wikis and search engines is beneficial to the understanding of this text.

Wilson Wong  
*University of Western Australia, Australia*

Wei Liu  
*University of Western Australia, Australia*

Mohammed Bennamoun  
*University of Western Australia, Australia*
Acknowledgment

This book would not have been possible without the efforts put in by the Editorial Advisory Board members and all contributing authors. In particular, a special thank you goes to the following individuals for their help in reviewing the submissions received for this book.

Christian Andrich, Graz University of Technology, Austria
Gabor Melli, PredictionWorks Inc, USA
Hongbo Deng, University of Illinois at Urbana-Champaign, USA
Johann Mitloehner, Vienna University of Economics and Business, Austria
Josef Moser, Graz University of Technology, Austria
Karin Verspoor, University of Colorado Denver, USA
Lipika Dey, Tata Consultancy Services, India
Martijn Schuemie, University Medical Center Rotterdam, the Netherlands
Sunam Kim, University of Melbourne, Australia
Yves Lussier, University of Chicago, USA
Jianbin Huang, Xidian University, China

The editors would also like to thank Mr. Joel Gamon, the Development Editor at IGI Global for his continuous support and advice.