Foreword

"To live effectively is to live with adequate information." Norbert Weiner, The Human Use of Human Beings (1954).

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

Ian Witten *is Professor of Computer Science at the University of Waikato in New Zealand where he directs the New Zealand Digital Library research project. His research interests include language learning, information retrieval, and machine learning. He has published widely, including several books, such as Managing Gigabytes (1999), Data Mining (2005), Web Dragons (2007), and How to Build a Digital Library (2003). He is a Fellow of the ACM and of the Royal Society of New Zealand. He received the 2004 IFIP Namur Award, a biennial honour accorded for "outstanding contribution with international impact to the awareness of social implications of information and communication technology" and (with the rest of the Weka team) the 2005 SIGKDD Service Award for "an outstanding contribution to the data mining field." In 2006, he received the Royal Society of New Zealand Hector Medal for "an outstanding contribution to the advancement of the mathematical and information sciences," and in 2010, was officially inaugurated as a "World Class New Zealander" in Research, Science, and Technology.*