
Preface

About the book

The success of the information society

The rapid progress of the “information society” in the past decade has been made possible by the removal of many technical barriers. Producing, storing, and transporting information in large quantities are no longer significant problems.

Producing on-line, digitized information is no longer a problem. Ever more of our commercial, scientific and personal information exchanges happen on-line in digital form. In the professional domain, near 100% of all office documents are produced in digital form (even if afterwards they are distributed in paper form), large parts of the scientific discourse are now taking place in digital form (with physics, computer science and astronomy taking a leading role). In the public domain, newspapers are available on-line, an increasing number of radio and television stations offer their material on-line in streaming form and e-government is an important theme for public administration. Even in the personal area, information is rapidly moving on-line: sales of digital cameras are now higher than for analogue cameras, e-mail and on-line chat have become important channels for maintaining social relations and for personal entertainment the digital DVD is rapidly replacing the analogue video tape. Compact disk (itself already digital) is under serious pressure from on-line music in MP3 format from a variety of sources. In short: production of on-line information is now the norm in virtually all areas of our life.

Storing such information in the required volumes is also no longer a problem. The drive of my laptop has truly become an on-line archive, both professionally and personally. As my professional archive, it stores the sources of around 100 scientific papers I have written (and the full sources of three books), all the Master theses of the dozens of students I have supervised, the

slides for countless presentations I have given, all the e-mail I have sent in the past 10 years plus all the e-mail I received in that period that I deemed worth keeping. But it also acts as a personal archive: my laptop holds all my favorite music, all the digital photographs I have ever taken, drawings and songs by my children, all my bank transfers of the past 10 years, and all my tax filings of the past 8 years. All of these data easily fit in a few tens of gigabytes, and occupy only a part of the storage capacity of my laptop.

Transport. Once we have created and stored our information on-line in digital form, it is also possible to move the information around in almost unlimited fashion: The Internet has solved most wide-area networking problems with its nearly universally supported TCP/IP protocol and its DNS host-addressing scheme. This global connectivity is now routinely available not only in offices, but also in households. Connectivity is also no longer a problem: a rapidly increasing percentage of households is on 1 Mbit/sec permanent connectivity, and connectivity at the workplace is typically at much higher bandwidth still.

The remaining problems

Given these nearly solved problems on production, storage and transport of information, what are the main remaining problems, if any? In an ironic way, it is exactly the above solutions that have created the most urgent remaining problems:

- *Information finding.* The large-scale and near-universal availability as a consequence of the successful technology mentioned above is as much a curse as it is a blessing. The more information is available, the harder it is to locate any particular piece of it.
- *Information integration.* Even when it is possible to find any particular piece of information, it is very hard to combine this information with any other piece of information we may already possess.

Typically, information is only meaningful in the context of other information, but most mechanisms we have available for publishing, locating and retrieving information deal with single, isolated instances of information, at the grain size of a document, a Web page or a diagram, and do not help us at all in integrating this information into what we already know.

Together, we call this problem with information finding and information integration the problem of *information sharing*. This general problem of information sharing occurs at many different levels, ranging from the overcrowded hard disk of our own PC, to knowledge-management problems in organizations, and to the sea of unstructured information on the World Wide Web.

The main thesis of this book is that the problem of information sharing (i.e. finding pieces of information and meaningfully relating them with other pieces) is only solvable by giving the computer better access to the *semantics* of the information. Thus, for a document, we do not only need to store such obvious metadata as its author, title, creation date, etc, but we must also make available in a machine-accessible way the important concepts that are discussed in the document, the relation of these concepts with those in other documents, relating these concepts to general background knowledge, etc. Similarly, for digital images, we would not only want to store format and size, but also that it is a satellite image of a specific area of land, where that area is located (e.g. by referencing a vocabulary of geographic locations), etc.

If computers had access to such *metadata* about the information items, they would be able to support us in finding relevant items, and in combining multiple items into a coherent answer to our questions. In this book we discuss active research on exactly this topic:

- how can the semantics of our information items be made available in a machine-accessible form?
- how can such metadata be exploited in retrieving and integrating information?

Of course it is crucial that the intended meaning of the metadata is shared between the different parties involved (e.g. those creating the metadata, and those using it). It is here that *ontologies* play a crucial role: shared formalized models of a particular domain, whose intended semantics is both shared between different parties and machine-interpretable (because it is “formalized”).

It has been argued that ontologies are a key technology for resolving the open problem of meaningful information sharing. However, most approaches rely on the existence of well-established data structures that can be used to analyze and exchange information. This book investigates ontology-based approaches for resolving semantic heterogeneity *weakly* structured environments, and in particular the World Wide Web. In doing this, we have to provide solutions for the following problems that arise from the nature of the Web:

Missing conceptual models: On the Web, we have no access to the conceptual model of an information source or the resulting logical data model. This lack of structure makes it difficult to refer to the context of information items, which is necessary for stating context transformation rules.

Unclear system boundaries: On the Web, it is not possible to clearly determine which information has to be taken into account, because information sources are added, removed or changed frequently. Therefore, we cannot rely on a fixed set of context-transformation rules.

Heterogeneous representations: On the Web, we can also not assume that ontologies are represented in a uniform way, because different representations are being used. This means that we also have to perform an integration on the ontology level.

Addressing these problems, this book contributes to a framework for ontology-based information sharing in weakly structured environments such as the *Semantic Web*.

Intended readership

This book is describing state-of-the-art research on these questions. As such the book is of potential interest for practitioners and applied researchers in the area of information systems, database technology and the Semantic Web.

For practitioners in areas such as e-commerce (exchange of product knowledge) and knowledge management (in particular in large and distributed enterprises), the book provides decision support for the use of novel technologies, information about potential problems and guidelines for the successful application of existing technologies.

The book draws on a large number of techniques from very different areas, such as terminological reasoning, inductive logic programming and query rewriting. To researchers in these different areas, the book provides evidence for the usefulness of various techniques from these different areas.

Organization of the Book

The topic of information sharing is a rather general one that stands for many different problems and technologies. In this book we try to give an overview of some of the most relevant technologies, restricting ourselves to the ideas and the technologies of the so-called "Semantic Web". Consequently, topics like ontologies, content metadata and reasoning about conceptual knowledge re-occur at many different places. Different methods for creating, maintaining and using ontologies and metadata are presented in the different chapters. Some of these technologies build upon each other, others are rather independent, but still contribute to the overall picture of technologies for information sharing on the Semantic Web. We tried to reflect this dependency in the overall organization of the book that is presented in the following.

The book is organized into four main parts.

Part I

introduces the general problem of information sharing and the need for explicit representations of information semantics in order to share information in a meaningful way. Further, it introduces the notion of ontology as a way of representing information semantics that has proven its value in different application domains. We also introduce the Web Ontology Language OWL as a standard for representing ontologies on the Semantic Web.

Part II

covers the creation of explicit representations of the information semantics. This includes the development of ontology encoded in OWL based on a given information sharing problem and the mostly automatic annotation of information sources with metadata that uses terms from ontologies to describe the content of an information source. We describe the basic methods for creating ontologies and metadata and describe experiments with real data and integration problems.

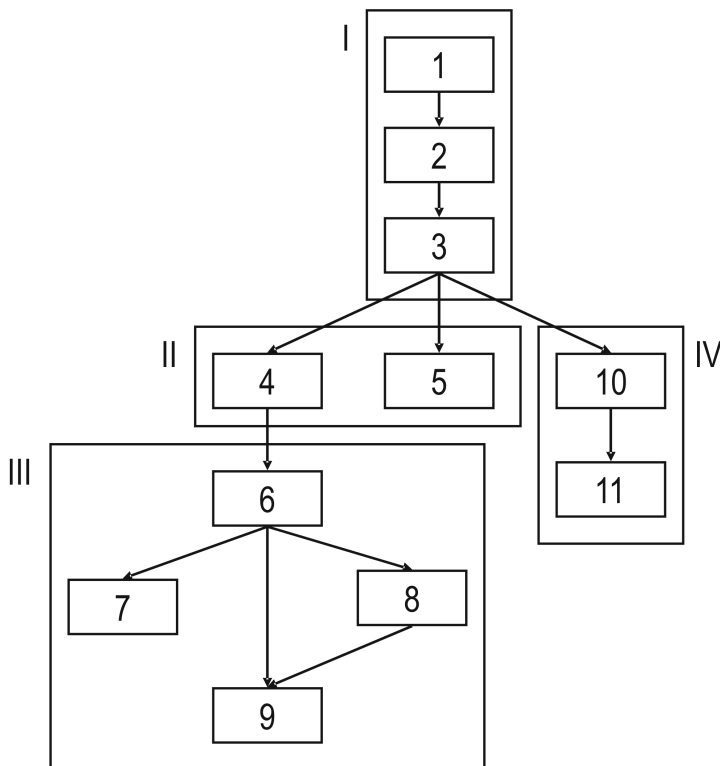
Part III

describes the use of the representational infrastructure created using the methods described in Part II for the purpose of information sharing. We discuss the semantic integration of terminologies used by different information sources and the integrated retrieval of information from multiple sources based on the result of the integration. Special attention is paid to the use of conjunctive queries that contain terms from ontologies. After discussing basic notions, we report the use of Semantic Web technologies for retrieving statistical information that revealed the need to take spatial relevance into account. We summarize with a description of the functionality of existing systems for information sharing and explain how the different aspects discussed in this part of the book are implemented in these systems.

Part IV

takes us back to some more fundamental questions concerning the use of ontologies for information sharing in a distributed environment such as the Semantic Web. In particular, we re-consider situations where the ontology itself is distributed across the Web. We extend the import mechanism of the Web Ontology Language by introducing the notion of modular ontology. We define a non-standard semantics for modular ontologies and compare the expressiveness of the model with OWL. We study the evolution of a modular ontology, in particular the impact of changes in a modular ontology, characterize changes according to their impact on other modules and define an update strategy that guarantees consistency of the overall model.

The drawing below illustrates the dependency between the different sections of the book. It is meant to guide readers only interested in particular aspects of information sharing. The first three chapters contain the motivation for the work and the introduction of central notions and representations such as ontologies and the Web Ontology Language OWL. All other parts of the book make use of these basic notions and can therefore only be completely understood after reading Chap. 1 to 3. Readers already familiar with Semantic Web technology, in particular ontologies and OWL might want to skip this part and only use it as a reference. After having read part I, the reader can decide to continue with part II or IV depending on the preferred focus.



Acknowledgements

Some of the content of this book has previously been published by organizations that are not part of the Springer Group. We thank these organizations for the kind permission to use the material for this book, in particular

- AAAI Press
- Elsevier

- IDEA Group Publishing

Some of the material reported here is the result of joint work with colleagues not mentioned as authors. We would like to thank the following persons for the fruitful cooperation in the past that led to the results reported in this book, as well as their permission to use material of joint papers and tutorials:

- Grigoris Antoniou, ITC-FORTH, Greece (Chap. 1)
- Fausto Giunchiglia, DIT, University of Trento, Italy (Chap. 6)
- Jens Hartmann, AIFB, University of Karlsruhe, Germany (Chap. 5)
- Catholijn Jonker, Vrije Universiteit Amsterdam, the Netherlands (Chap. 7)
- Michel Klein, Vrije Universiteit Amsterdam, the Netherlands (Chap. 10 and 11)
- Eduardo Mena, University of Zaragoza, Spain (Chap. 9)
- Christoph Schlieder, University of Bamberg, Germany (Chap. 8)
- Tim Verwaart, LEI Wageningen, the Netherlands (Chap. 7)
- Ubbo Visser, TZI, University of Bremen, Germany (Chap. 1, 2, 8 and 9)
- Thomas Voegelé, TZI, University of Bremen, Germany (Chap. 1, 2, 8 and 9)
- Holger Wache, TZI, University of Bremen, Germany (Chap. 2, 6 and 9)

Some of the work has been supported by the European Union under contracts IST-2001-33052 (WonderWeb) and IST-2001-34103 (SWAP). A significant part of the work has been carried out by the first author during his appointment at the Artificial Intelligence Group (Prof. Herzog) at the University of Bremen, Germany.

Amsterdam,
January 2004

Heiner Stuckenschmidt
Frank van Harmelen

Information Sharing on the Semantic Web

Stuckenschmidt, H.; Harmelen, F. van

2005, XIX, 276 p., Hardcover

ISBN: 978-3-540-20594-4