Terminological Ontologies

SEMANTIC WEB AND BEYOND
Computing for Human Experience

Series Editors:

Ramesh Jain
University of California, Irvine
http://ngs.ics.uci.edu/

Amit Sheth
Wright State University
http://knoesis.wright.edu/amit/

As computing becomes ubiquitous and pervasive, computing is increasingly becoming an extension of human, modifying or enhancing human experience. Today's car reacts to human perception of danger with a series of computers participating in how to handle the vehicle for human command and environmental conditions. Proliferating sensors help with observations, decision making as well as sensory modifications. The emergent semantic web will lead to machine understanding of data and help exploit heterogeneous, multi-source digital media. Emerging applications in situation monitoring and entertainment applications are resulting in development of experiential environments.

SEMANTIC WEB AND BEYOND
Computing for Human Experience
addresses the following goals:

- brings together forward looking research and technology that will shape our world more intimately than ever before as computing becomes an extension of human experience;
- covers all aspects of computing that is very closely tied to human perception, understanding and experience;
- brings together computing that deal with semantics, perception and experience;
- serves as the platform for exchange of both practical technologies and far reaching research.

Additional information about this series can be obtained from
http://www.springer.com/series/7056

Javier Lacasta • Javier Nogueras-Iso
Francisco Javier Zarazaga-Soria

Terminological Ontologies

Design, Management and Practical Applications

Springer

Javier Lacasta
University of Zaragoza
Maria de Luna 1
50018 Zaragoza
Spain
jlacasta@unizar.es

Javier Nogueras-Iso
University of Zaragoza
Maria de Luna 1
50018 Zaragoza
Spain
jnog@unizar.es

Francisco Javier Zarazaga-Soria
University of Zaragoza
Maria de Luna 1
50018 Zaragoza
Spain
javy@unizar.es

ISSN 1559-7474
ISBN 978-1-4419-6980-4 e-ISBN 978-1-4419-6981-1
DOI 10.1007/978-1-4419-6981-1
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2010931829

To María del Carmen, Aurelio and Elena.
To Ascensión, Fermín and María Jesús.
To Esther, Victor and Silvia.
41.6839, -0.8892 —— 41.6677, -0.8788

Preface

Information infrastructures are integrated solutions based on the fusion of information and communication technologies. An information infrastructure is defined as an advanced, seamless web of public and private communications networks, interactive services, interoperable hardware and software, computers, databases, and consumer electronics to make available vast amounts of information. The term started to be commonly used after the launching of the US plan for National Information Infrastructures [195]. Since then, the term has been widely used to describe national and global communication networks like the Internet and more specialized solutions for communications within specific business sectors. For example, following the US path, the European Union published some years later its own plan for the creation of European information infrastructures [46].

Information retrieval is a basic functionality in any information infrastructure. Information retrieval deals with the representation, storage, organization, and access to information items [13]. It consists in determining which documents of a collection are relevant to the user information request. The primary goal of an information retrieval system is to retrieve all the documents that are relevant to the user information need while retrieving as few non-relevant documents as possible. To do so, it has to be able to extract syntactic and semantic information from the documents, and use this information to rank the documents according to the degree of match with respect to the user information need. However, the interpretation of the user need is not an easy task. It is limited by the expressivity of the user query language and by the inherent ambiguity and terminological dispersion of the written text.

An information infrastructure requires an efficient and effective information retrieval system to provide the users with access to the items stored in the infrastructure. It does not really matter how much information about a subject an infrastructure contains; if it is not possible to find it, it is useless. Therefore, it is important to distinguish between an information retrieval and a data retrieval process. While data retrieval systems are focused on determining which records stored in a catalog system contain the words specified in the user query, information retrieval ones are more concerned with obtaining information about a subject or topic than retrieving the data which satisfies exactly a given query. Data retrieval techniques are

applicable to systems with well structured data where returning a single erroneous item means a total failure. However, in systems working with natural language text, which is not always well structured and could be semantically ambiguous, information retrieval systems could be a better option if some inaccuracies and small errors are acceptable.

An information infrastructure is composed of several services and components that have to interact to provide the desired functionality. If each component uses a different set of interfaces and formats, the interoperability between them becomes a difficult task. The use of standards is of great help to solve syntactic interoperability problems establishing a common way to access to information (they provide a common syntax). However, syntactic interoperability is not enough for information retrieval. Natural language terms used in classification, indexing and querying contain semantic relations between them (e.g., synonymy, polysemy, homonymy, meronymy, hyponymy, lexical variants, or misspellings) may make difficult the creation of effective search services.

In order to increase semantic interoperability in search systems, libraries, museums, and archives have traditionally used controlled vocabularies (list of terms about a certain subject) to describe resources, reducing in that way the possible terms used in classification and search to the selected ones. Their use increases the homogeneity in the descriptions, simplifies the query process and improves the results. Controlled vocabularies are used in classification steps to describe (and index) the resources. In the search components, they provide the user with the appropriate terminology for constructing queries. And in information browsing, they are used to provide a browsing structure through the resources based on the selected vocabulary. The selection of an appropriate vocabulary represents nevertheless an important challenge [74], it has to be adapted to the collection requirements avoiding terms irrelevant for the desired context.

Having in mind the increase of terminological precision, the use of simple controlled vocabularies has been progressively displaced by the use of more sophisticated knowledge models. This tendency has been greatly increased in the last years with the impact of Internet and the Semantic Web. The knowledge models stored in paper (taxonomies, thesauri) by libraries and other institutions have been computerized and transformed into more formal ontology models to provide a higher level of semantics. The term ontology is used in information systems and in knowledge representation systems to denote a knowledge model, which represents a particular domain of interest. A body of formally represented knowledge is based on a conceptualization: the objects, concepts, and other entities that are assumed to exist in some area of interest and the relationships that hold among them. And an ontology provides "an explicit formal specification of a shared conceptualization" [69]. Some ontology types are classification schemes that organize materials at a general level (such as books on a shelf), subject headings that provide more detailed access, authority files that control variant versions of key information (such as geographic names and personal names), or semantic networks and formal ontologies that provide a complete set of formally defined relations. Depending on the formalism level, Sowa [178] distinguishes two main classes of ontologies: terminological

(also called lexical) and axiomatized (also called formal). Terminological ontologies are not fully specified by axioms and definitions and the relations are limited to subtype/supertype or part/whole relations. On the other hand, axiomatized ontology concepts and relations have associated axioms and definitions that are stated in logic or in some computer-oriented language that can be automatically translated to logic.

Nowadays, ontologies play an important role in the information retrieval context. Usually, the retrieval of resources through an information retrieval system implies successive iterations of resource discovery, followed by resource evaluation, and a final stage of access (direct as a data set, or indirect via a data access service) and exploitation of the resource. In all these stages, different ontology models can help to improve the produced results.

As concerns **resource discovery**, some of the most remarkable problems that affect the interoperability and cooperation of discovery systems are metadata schema heterogeneity and content heterogeneity [156].

In order to facilitate discovery and access, the content of a collection is summarized into small descriptions, usually called metadata (data about the data), which can be either introduced manually or automatically generated (index terms automatically extracted from a collection of documents). Most Digital libraries define their structured metadata in accordance to recognized standards such as MARC21 [149] or the Dublin Core Metadata Element Set [85] (proposed by the Dublin Core Metadata Initiative[1]) but other models can be used. This heterogeneity makes difficult the integration of collections using different metadata models. In this context, given that a metadata schema is a model that contains a set of concepts with properties and relations to other concepts, their structure can be modeled as a formal ontology, where metadata records are instances of this ontology [17]. This kind of formal ontologies may be used to profile the metadata needs of a specific resource and its relationships with metadata of other related resources, or to provide interoperability across metadata schemas.

On the other hand, metadata try to describe in an accurate way information resources to enhance information retrieval, but this improvement depends greatly on the quality of metadata content. Even in the same collection, the content of each metadata can be quite heterogeneous. Here, terminological ontologies facilitate classification of resources and information retrieval. One way to enforce the quality is the use of selected terminology for some metadata fields in the form of lexical ontologies. These ontologies are used to describe contents but also allow computer systems to reason about them. This role of terminological ontologies is even more significant in the case of developing multilingual systems because they can provide the translations of the terms used for classification to all the required languages.

Regarding **resource evaluation**, an information retrieval system and must provide enough means to visualize the data appropriately. In this scenario, one could consider multilinguality and other specific issues related with the type of data stored.

In the case of viewing metadata in a specific language required by the user, one may face the problem of having to translate it. Once again, formal ontologies and

[1] http://www.dublincore.org

terminological ontologies may facilitate the work in two important aspects. Firstly, a formal ontology may provide the labels, in the appropriate language, for the elements of the metadata schema. Secondly, terminological ontologies may be used in the task of automatic translation of metadata to increase accuracy of translations.

Depending on the processed data, other issues where the use of ontologies helps may raise. For example, in a map repository, developers must consider the internationalization of legends and the display of internationalized attribute information.

Finally, the **resource access** step may benefit as well from the use of ontologies to facilitate data sharing and system development. Once again, formal ontologies help to define the meaning of features or a resource and they can provide a "common basis" for semantic mapping, e.g. to find similarity between two features that represent the same object but that have been defined using different language representations.

From the different types of ontologies, terminological ones are the most commonly used in every aspect of an information retrieval process. Their uses range from classification to query construction and results visualization. However, if all these different models are created, used and managed independently, the complexity of the system increases in a great deal. This book focuses on providing a coherent framework for the integration of terminological ontologies in an information retrieval system, with the objective of facilitating its creation, management, and use for the different components requiring it. The integration problems that have been faced can be divided into three main general categories: representation, acquisition and access:

- Related to the **representation** problem, it is common to find that each organization has created a new ontology using an ad-hoc representation format, which is only useful in its specific context. This has led to a big heterogeneity of representation models that increases the difficulty and the cost of integrating the models into a homogeneous system. In this context, a single and homogeneous representation mechanism for terminological models is vital to provide uniform ontology models to the components that require them. An additional problem is the need to provide a single and homogeneous access to different data collections classified with different terminological ontologies. For example, when integrating data from different countries classified according to different terminological models in different languages. To provide a homogeneous access to the resources, the used ontologies have to be related to be able to identify equivalences and obtain complete results. The process of matching ontologies (called ontology alignment) is difficult and costly but other collections using the same terminological models can reuse it. Therefore, in a similar way as it is required a representation format for individual terminological models, it is required another one for storing the mappings between them.
- Regarding the ontology **acquisition** problem, the needed ontologies have to be obtained or created, and adapted to integrate them in the required systems. However, this is not an easy task. On the one hand, the heterogeneity in the creation of terminological models limits their reusability in contexts different from the original ones. Therefore, even if a suitable terminology is found, it has to be

transformed to facilitate its integration with the rest used in the system. This requirement involves additional integration issues due to the need of a different transformation process for each required ontology. On the other hand, the creation of a new one from scratch is very costly in time and resources. In this context, to reduce the development effort, it is useful to reuse sections of other ontology models that contain suitable terminology. An additional issue that has been taken into account is the overlapping of the acquired models. Here, the common elements of different models have to be properly managed to avoid classification problems.

- Finally, with respect to the **access** problem, the applicability of these models in a wide range of application domains has led to the creation of a great variety of terminological ontologies with very different levels of specificity, language coverage (i.e., from monolingual list of terms to multilingual thesauri covering dozens of languages), formalization (i.e., from simple glossaries to well-structured thesauri), or size (e.g., AGROVOC thesaurus [126] contains more than 16,000 concepts). Additionally, it is important to note that they are distributed to the public through ad-hoc services created for each institution providing them. This ad-hoc distribution is not appropriate in an information infrastructure where it is required to provide the ontologies to all components in a simple and common way. In this context, it is needed a coordinated view of the ontologies that can only be obtained through a homogeneous management and access not dependent of the original providers.

In order to solve these specific problems, this book describes a homogeneous solution for each of these discovery scenarios. These solutions are interrelated in such a way that they can be combined to facilitate all the steps required to integrate a new terminological ontology into an information retrieval system.

1. In order to deal with the representation issues, the existent representation formats for terminological models have been analyzed. From them, the most appropriate has been selected, extending it to cover those information requirements that were not fulfilled in the original format model. A similar work has been done with respect to the representation of mappings between different terminological models. In this case, given that no suitable format exists, a new one based on textual recommendations indicated in the terminological ontology standards has been designed.
2. With respect to the acquisition issue, each problem described has required a different approach as part of a global transformation process. First, a general transformation process is proposed to harmonize the way a terminological ontology is converted to the selected representation format. The format allows defining the structure of the source and destination models and simplifies the definition of relations between them. The proposed architectural pattern helps to reuse the common elements of the different transformations. Secondly, to simplify the construction of a new ontology, a process that uses a set of ontologies as base and combines them into a new model is described. To focus the result into the desired domain, the process limits the content of the new ontology pruning the

non-relevant concepts. Finally, with the objective to increase the formalization of the models when required, a process that helps in the identification of the existent *is-a* relationships has been developed.

3. In the ontology management context, it has been identified the need for an efficient and common ontology management service to filter and select the most appropriate ontology for each specific context. However, before the creation of these services, the design of a common repository is proposed to store all the required terminological ontologies. On top of this repository, the design of an efficient editor and other GUI widgets is proposed to facilitate the annotation and the update of terminological ontologies. Additionally, a centralized ontology service, called Web Ontology Service (WOS), which enables uniform management of terminological ontologies (including discovery services) has been developed to provide access to terminological ontologies via Web services. To provide a full integration with the rest of components of a typical information retrieval system, it follows and extends standard interfaces used by the Semantic Web community.

This book consists of seven chapters describing in detail the integration problems and the proposed solutions. The content of these chapters is organized as follows:

- Chapter 1 reviews the types of ontologies and the techniques used for establishing alignments between them. The concepts, ideas and techniques described are used along the entire book.
- Chapter 2 focuses on the problems of representation of terminological ontologies. It starts analyzing the problematic of representation of simple models, and then follows with the issues related to representing the relationships between overlapping models.
- Chapter 3 analyzes the issues related to the creation of terminological ontologies using data from different sources such as data corpora, dictionaries, schemata and other knowledge models.
- Chapter 4 describes the issues related to the lack of formalism in terminological models and how to increase it in order to provide additional semantic functionality.
- Chapter 5 analyzes the way to provide access to terminological models. It describes the structure of a terminological ontology repository, a tool for managing and editing the ontologies, and a web service for providing access to them.
- Chapter 6 describes how the components proposed in chapter 5 can be integrated into an information retrieval system. As a result of this integration we present: tools for creating metadata facilitating the management of terminological models for classification; search clients able to use the stored ontologies to improve the search results; and browsing systems providing access to the resources on the basis of the structure of a terminological ontology.
- Chapter 7 contains some concluding remarks and an outlook of future areas of work.

Zaragoza,
April 2010

Javier Lacasta
Javier Nogueras-Iso
F. Javier Zarazaga-Soria

Acknowledgements

There are many people to whom we are grateful for their support during the evolution of this book.

First of all, we would like to thank to the members and friends of the Advanced Information Systems Laboratory (IAAA) of the University of Zaragoza, making a especial mention to Pedro R. Muro-Medrano, Rubén Béjar, and F. Javier Lopez-Pellicer by their comments and suggestions that have been so valuable for improving this book. We also have to include other members and ex-members from the staff of IAAA and GeoSpatiumLab S.L. that have provided us with the required technical support and advice, specially to Juanjo, Jesús, Mariano, Rodolfo, Covadonga, Christian, Aneta, and Miguel Ángel. Besides, we cannot forget the support of our colleagues at the Computer Science and Systems Engineering Department, specially Jose Ángel Bañares, Joaquín Ezpeleta and Pedro Álvarez.

Finally, we are absolutely grateful to our families and friends for all their patience, support and love. Much time have been stolen from our personal lives for the creation of this text. Undoubtedly, without their generous understanding this work would never have come into existence.

Contents