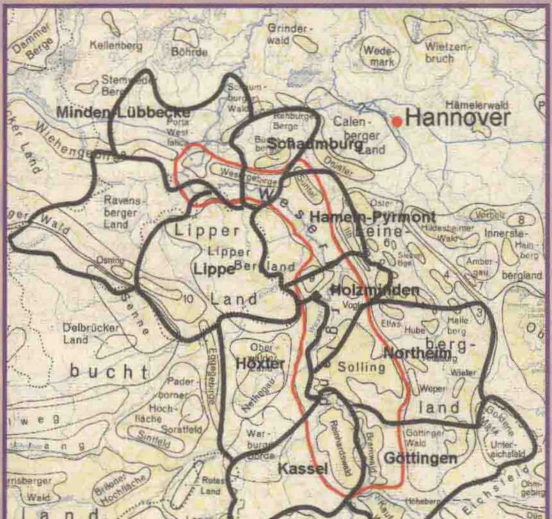
Ubbo Visser

LNAI 3159

# Intelligent Information Integration for the Semantic Web



Springer

Ubbo Visser

# Intelligent Information Integration for the Semantic Web

Springer

Author

Ubbo Visser
University of Bremen
Center for Computing Technologies (TZI)
Universitätsallee 21-23, 28359 Bremen, Germany
E-mail: visser@tzi.de

Library of Congress Control Number: 2004110975

CR Subject Classification (1998): I.2, H.2, H.3, H.4

ISSN 0302-9743
ISBN 3-540-22993-0 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springeronline.com

Typesetting: Camera-ready by author, data conversion by Olgun Computergrafik
Printed on acid-free paper SPIN: 11318583 06/3142 5 4 3 2 1 0

Dedicated to my family Susan and Jannes as well as my parents
who always gave me support in the rough times...

# Foreword

The Semantic Web offers new options for information processes. Dr. Visser is dealing with two core issues in this area: the integration of data on the semantic level and the problem of spatio-temporal representation and reasoning. He tackles existing research problems within the field of geographic information systems (GIS), the solutions of which are essential for an improved functionality of applications that make use of the Semantic Web (e.g., for heterogeneous digital maps). In addition, they are of fundamental significance for information sciences as such.

In an introductory overview of this field of research, he motivates the necessity for formal metadata for unstructured information in the World Wide Web. Without metadata, an efficient search on a semantic level will turn out to be impossible, above all if it is not only applied to a terminological level but also to spatial-temporal knowledge. In this context, the task of information integration is divided into syntactic, structural, and semantic integration, the last class by far the most difficult, above all with respect to contextual semantic heterogeneities.

A current overview of the state of the art in the field of information integration follows. Emphasis is put particularly on the representation of spatial and temporal aspects including the corresponding inference mechanisms, and also the special requirements on the Open GIS Consortium.

An approach is presented integrating information sources and providing temporal and spatial query mechanisms for GIS, i.e., the BUSTER system developed at the Center for Computing Technologies (TZI) which was defined according to the following requirements:

- Intelligent search
- Integration and/or translation of the data found
- Search and relevance for spatial terms or concepts
- Search and relevance for temporal terms

While distinguishing between the query phase and the acquisition phase, the above serves as the basis for the concept of the systems architecture. The representation of semantic properties requires descriptions for metadata: this is where the introduced methods of the Dublin Core are considered, and it is demonstrated that the elements defined there do not meet with the requirements and consequently have to be extended.

Furthermore, important problems of terminological representation, terminological reasoning, and semantic translation are treated extensively. Again, the definition of requirements and a literature survey on the existing approaches (ontologies, description logics, inference components, and semantic translation) sets the scope. The chapter concludes with a comprehensive real-world example of semantic translation between GIS catalogue systems using ATKIS (official German catalogue) and CORINE (official European catalogue) illustrating the valuable functions of BUSTER.

Subsequently, the author attacks the core problems of spatial representation and spatial reasoning. The requirements list intuitive spatial denominations, place-names, gazetteers, and footprints, and he concludes that existing results are not expressive enough to enable the desired functionalities. Consequently, an overview of the formalisms of place-name structures is given which is based on tessellations and allows for an elegant solution of the problem through a representation with connection graphs, including an evaluation of spatial relevance. The theoretical background is explained using a well-illustrated example.

Finally, the requirements for temporal representations and the corresponding inference mechanisms are discussed. A qualitative calculus is developed which makes it possible to cover the temporal aspects which are also of importance to Semantic Web applications.

After the discussion of the set of requirements for an intelligent query system, the state of the BUSTER implementation is discussed. In a comprehensive demonstration of the system, terminological, spatial, and temporal queries, and some of their combinations are described.

An outlook on future research questions follows. In the bibliography, a good overview is given on the current state of the research questions dealt with.

This book combines in an exemplary manner the theoretical aspects of a combination of intelligent conceptual and spatio-temporal queries of heterogeneous information systems. Throughout the book, examples are provided using GIS functionality. However, the theoretical concept and the prototypical system are more general. The ideas can be applied to other application domains and have been demonstrated and tested, e.g., in the electronics and tourist domains. This demonstrates well that the approaches worked out are useful for practical applications – a valuable benefit for those readers who are looking for actual research results in the important areas of data transformation, the semantic representation of spatial and/or temporal relations, and for applications of metadata.

Bremen, May 2004 Otthein Herzog

# Preface

When I first had the idea about the automatical transformation of data sets, which we now refer to as semantic translation, many of my colleagues were sceptical. I had to convince them, and when I showed up with a real-world example (ATKIS-CORINE) we founded the BUSTER group. This was in early 1999.

Since then, many people were involved in this project who helped with their critical questions, valuable suggestions, and ideas on how to develop the prototype. Two important people behind the early stages of the BUSTER idea are Heiner Stuckenschmidt and Holger Wache. I would like to thank them for their overview, their theoretical contributions, and their cooperation. I really enjoyed working with them and we hopefully will be able to do some joint work in the future again.

Thomas Vögele played an important role in the work that has been done around the spatial part of the system. His contributions in this area are crucial and we had fruitful discussions about the representation and reasoning components of the BUSTER system. At this point, I also would like to thank Christoph Schlieder, who gave me a thorough insight into the qualitative spatial representations and always contributed his ideas to our objectives. Some of them are now implemented in the BUSTER prototype.

The development and implementation of the system would not have been possible without people who are dedicated to programming. Most of the Master's students involved in our project were working on it for quite a long time. Sebastian Hübner, Gerhard Schuster, Ryco Meyer, and Carsten Krüwel were amongst the first "generation". I would like to thank them for their programming skills and patience when I asked them to have something ready as soon as possible. Sebastian Hübner now plays an important role in our project. Without him, the new temporal part of our system would be non-existent.

Bremen,
April 2004

*Ubbo Visser*

# Table of Contents

## Part I Introduction and Related Work

## Part II The Buster Approach for Terminological, Spatial, and Temporal Representation and Reasoning

## Part III Implementation, Conclusion, and Future Work

# Part I

# Introduction and Related Work

# 1

# Introduction

The Internet has provided us with a new dimension in terms of seeking and retrieving information for our various needs. Who would have thought about the vast amount of data that is currently available electronically ten years ago? When we look back and think about what made the Internet a success we think about physical networks, fast servers, and comfortable browsers, just to name a few. What one might not think about, a simple but important issue is the first version of HTML. This language allowed people to share their information in a simple but effective way. All of a sudden, people were able to define a HTML document and put their information piece on the Web. The given language was sloppy and almost anybody with a small amount of knowledge about syntax or simple programming use could define a web page. Even when language items such as end-tags or closing brackets were forgotten, the browser did the work and delivered the content without returning syntax errors. We believe this to be a crucial point when considering the success story of the Internet: give the people a simple but effective tool with the freedom to provide their information.

Providing information is one thing, searching and retrieving information is at least as important. Early browsers or search engines offered the opportunity to search for specific keywords, mostly searching for strings. The user was prompted with results in a rather simple way and had to choose the required information manually. The more data were added to the Web, the harder the search for information became. The latest versions of search engines such as Google provide a far more advanced search based on statistical evidences or smart context comparisons and rank the results accordingly. However, the users still have to choose the information they are interested in more or less manually.

Being able to provide data in a rather unstructured or semi-structured way is part of the problems with automatic information retrieval. This is the situation behind the activities of the W3C concerning the Semantic Web. The W3C defines the Semantic Web on their Web page as:

> "The Semantic Web is the abstract representation of data on the World Wide Web, based on the RDF standards and other standards to be defined. It is being developed by the W3C, in collaboration with a large number of researchers and industrial partners." [136][1]

The same page contains a definition of the Semantic Web that is of similar importance. This definition has been created by [8] and states

> "The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation." [136][2]

These definitions indicate the Web of tomorrow. If data have a well-defined meaning, engines will be able to intelligently seek, retrieve, and integrate information and generate new knowledge to answer complex queries.

The retrieval and integration of information is the focus of this paper. Before going into detail we would like to share some creative ideas, which can be a vision of what we can expect from the Semantic Web.

## 1.1 Semantic Web Vision

Bernes-Lee et al. [8] already gave us an insight of what we should be able to do with the help of data and engines working in the Web. In addition, the following can help to see where researchers want to arrive in the future. These ideas can be distinguished into four groups:

- **Short-term:** The following tasks are not far away from being solved or, are already solved to a certain extent.
  - *Being able to reply on an email via telephone call:* This requires communication abilities between a phone and an email client. Nowadays, the first solutions are available, however, vendors offer a complete solution with a phone and an email client that come in one package with more or less the same software. An example is the VoiceXML package from RoadNews[3]. The beauty of this point is that an arbitrary email client and an arbitrary phone can be used. The main subject is interoperability between address databases.
  - *Meaningful browsing support:* The idea behind this is that the browser is smart enough to detect the subject the user is looking for. If for instance, the user is looking for the program on television for a certain day on a web page, the browser could support the user by offering similar links to other web sites offering the same content.

[1] http://www.w3.org/2001/sw/, no pagination, verified on Oct 17, 2002.

[2] http://www.w3.org/2001/sw/, no pagination, verified on July 1st, 2003.

[3] http://www.roadnews.com, verified on July, 1st, 2003.

- **Mid-term:** These tasks are harder to solve and we believe that solutions will be available in the next few years.
  - *Planning appointments with colleagues by integrating diaries:* This is a problem already tackled by some researchers (e.g. [90]) and the first solutions are available. Pages can be parsed to elicit relevant information and through reference to published ontologies reasoning support, it is possible to provide user assistance. However, this task is not simple and many problems still have to be addressed. This task serves as one example of the ongoing Semantic Web Challenge (http://challenge.semanticweb.org).
  - *Context-aware applications:* Ubiquitous computing might serve as another keyword in this direction. Context-awareness (cf. [49]) has to deal with mobile computing, reduction of data, and useful abstraction (e.g., digital maps in an unknown city on a PDA).
  - *Giving restrictions for a trip and getting the schedule and the booking*: The scenario behind this is giving a computer the constraints for a vacation/trip. An agent is then supposed to check all the information available on the Web, including the local travel agencies and make the booking accordingly. Besides some severe technical problems, such as technical interoperability between agencies, we also have to deal with digital signatures and trust for the actual booking at this point. First approaches include modern travel portals such as DanCenter[4] where restrictions for a trip can be made and booking is also possible. This issue will be postponed for now.
- **Long-term:** Tasks in this group are again more difficult and the solutions might emerge only in the next decade.
  - *Information exchange between different devices:* Suppose, we are surfing the Web and see some movies we are interested in which will be shown on television during the next few days. Theoretically, we are able to directly take this information and program our VCR (e.g., WebTV[5]).
  - *Oral communication with the Semantic Web:* So far, plain commands can be given via speech software to a computer. This tasks goes even further: here, we think about the discussions of issues rather than plain commands. We also anticipate inferences and interaction.
  - *Lawn assistant:* Use satellite and weather information from the Web, background garden knowledge issued to program your personal lawn assistant.
- **Never:** Automatic fusion of large databases.

We can identify a number of difficult tasks that will most likely be difficult to solve. The automatic fusion of large databases is an example for this. On the other hand, we have already seen some solutions (or partly solutions) for

[4] http://www.dancenter.com, verified on July, 1st, 2003.

[5] http://about-the-web.com/shtml/WebTV.shtml, verified on June, 1st, 2003.

tasks that are grouped into short- and mid-term problems (e.g., integrating diaries). The following research topics can be identified with regard to theses ideas.

## 1.2 Research Topics

The research topics are as numerous as the problems. The number of areas discussed at the first two International Semantic Web Conferences in 2001/2002 [19, 60] can be seen as an indication of this. Some of the topics were: agents, information integration, mediation and storage, infrastructure and metadata, knowledge representation and reasoning, ontologies, and languages. These topics are more or less concerned with the development and implementation of new methods and technologies. Topics such as trust, growth and economic models, socio-cultural and collaborative aspects also belong to these general issues with regard to the Semantic Web and are concerned with other areas.

We will focus on some of the topics mentioned first: metadata and ontologies, or more general knowledge representation and reasoning with the help of annotated information sources. In general, we have to decide on an appropriate language to represent the knowledge we need. We have to bear in mind that this language has to be expressive enough to cover the necessary elements of the world we are modeling. On the other, hand we have to think about the people who are or will be using this language to represent and annotate their knowledge or information sources needed to be accessible via WWW. If we do not expect highly qualified knowledge engineers to do this job (which is unrealistic if we want to be successful with the Semantic Web) we need to compromise between the complexity and the simplicity of the language[6].

We will discuss how ontologies are used in the context of the Semantic Web in section 2. When we say 'ontology' we refer to Gruber's well-know definition [45], that an ontology is an explicit specification of a conceptualization. Please note that we do not focus on terminological ontologies only. The vision of the Semantic Web clearly reveals that also spatial information (e.g., location-based applications, spatial search) and temporal information (e.g., scheduling trips, booking vacations) will be needed. We will motivate our research interests with two important issues: firstly, how do we find information or better: can we improve nowadays search engines? Secondly, once we have found information, how do we integrate this information in our application? The next two sections give a brief overview about what has to be considered with regard to search and integration of information.

[6] This is an analogy to the growth of the "old" Internet. The simplicity of HTML was one of the keys for the success of the WWW. Almost everybody was able to create a simple Web page with some text and/or picture elements. There was no syntax check telling the user that there is a bracket open and he/she has to fix it. The browser showed a result and did forgive little mistakes. This sloppiness was important because it helped a vast amount of people (non-computer scientist) to use HTML.