Arno Scharl
Klaus Tochtermann (Eds.)

The Geospatial Web

How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society

Springer

Advanced Information and Knowledge Processing

Also in this series

Gregoris Mentzas, Dimitris Apostolou, Andreas Abecker and Ron Young
Knowledge Asset Management
1-85233-583-1

Michalis Vazirgiannis, Maria Halkidi and Dimitrios Gunopulos
Uncertainty Handling and Quality Assessment in Data Mining
1-85233-655-2

Asunción Gómez-Pérez, Mariano Fernández-López and Oscar Corcho
Ontological Engineering
1-85233-551-3

Arno Scharl (Ed.)
Environmental Online Communication
1-85233-783-4

Shichao Zhang, Chengqi Zhang and Xindong Wu
Knowledge Discovery in Multiple Databases
1-85233-703-6

Jason T.L. Wang, Mohammed J. Zaki, Hannu T.T. Toivonen and Dennis Shasha (Eds.)
Data Mining in Bioinformatics
1-85233-671-4

C.C. Ko, Ben M. Chen and Jianping Chen
Creating Web-based Laboratories
1-85233-837-7

Manuel Graña, Richard Duro, Alicia d'Anjou and Paul P. Wang (Eds.)
Information Processing with Evolutionary Algorithms
1-85233-886-0

Colin Fyfe
Hebbian Learning and Negative Feedback Networks
1-85233-883-0

Yun-Heh Chen-Burger and Dave Robertson
Automating Business Modelling
1-85233-835-0

Dirk Husmeier, Richard Dybowski and Stephen Roberts (Eds.)
Probabilistic Modeling in Bioinformatics and Medical Informatics
1-85233-778-8

Ajith Abraham, Lakhmi Jain and Robert Goldberg (Eds.)
Evolutionary Multiobjective Optimization
1-85233-787-7

K.C. Tan, E.F. Khor and T.H. Lee
Multiobjective Evolutionary Algorithms and Applications
1-85233-836-9

Nikhil R. Pal and Lakhmi Jain (Eds.)
Advanced Techniques in Knowledge Discovery and Data Mining
1-85233-867-9

Amit Konar and Lakhmi Jain
Cognitive Engineering
1-85233-975-6

Miroslav Kárný (Ed.)
Optimized Bayesian Dynamic Advising
1-85233-928-4

Yannis Manolopoulos, Alexandros Nanopoulos, Apostolos N. Papadopoulos
and Yannis Theodoridis
R-Trees: Theory and Applications
1-85233-977-2

Sanghamitra Bandyopadhyay, Ujjwal Maulik, Lawrence B. Holder and Diane J. Cook (Eds.)
Advanced Methods for Knowledge Discovery from Complex Data
1-85233-989-6

Marcus A. Maloof (Ed.)
Machine Learning and Data Mining for Computer Security
1-84628-029-X

Sifeng Liu and Yi Lin
Grey Information
1-85233-995-0

Vasile Palade, Cosmin Danut Bocaniala and Lakhmi Jain (Eds.)
Computational Intelligence in Fault Diagnosis
1-84628-343-4

Mitra Basu and Tin Kam Ho (Eds.)
Data Complexity in Pattern Recognition
1-84628-171-7

Samuel Pierre (Ed.)
E-Learning Networked Environments and Architectures
1-84628-351-5

Arno Scharl and Klaus Tochtermann (Eds.)

The Geospatial Web

How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society

Springer

Prof. Arno Scharl
Prof. Klaus Tochtermann
Know-Center & Graz University of Technology, Graz, Austria

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Control Number: 2007926113

AI&KP ISSN 1610-3947
ISBN-13: 978-1-84628-826-5
e-ISBN-13: 978-1-84628-827-2

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

Springer Science+Business Media
springer.com

Foreword

The most important attribute of geospatial platforms is their unique potential to aggregate a multitude of public and private geographic data sets, providing access to data from government agencies, industry and the general public. NASA and other organizations have a wealth of planetary science data – representing the output from thousands of satellites in earth-orbit, and from dozens of costly missions to other planets. Benefits derived from both the data and visual interfaces to access the data represent a significant return on investment for the public. Integrating geospatial data with semantic and collaborative Web technology multiplies the public benefits and represents the main focus of this book.

The user interfaces of geobrowsers are designed for the layperson, giving convenient access to all kinds of geographically referenced information. Geobrowsers hide the technical details related to finding, accessing and retrieving such information. The daunting challenge of the Geospatial Web is to seamlessly integrate and display vastly different information modes. Nowadays, it is not enough to simply display a map of some region; additional dynamic information modes need to be displayed and put into context – from weather sensor readings and live aerial video feeds to daily news updates, photo collections and video archives.

The open-source community plays a crucial role in driving the development of the Geospatial Web. Collaborative efforts have provided a large number of add-ons for popular platforms. In the case of NASA World Wind, several of these external modules have been integrated into the core system. Participants in open-source projects identify, track and resolve technical problems, suggest new features and source code modifications, and often provide high-resolution data sets and other types of user-generated content.

This book presents the state-of-the-art in geospatial Web technology. It gradually exposes the reader to the technical foundations of the Geospatial Web, and to new interface technologies and their implications for human-computer interaction. Several chapters deal with the semantic enrichment of electronic resources, a process that yields extensive archives of Web documents, multimedia data, individual user profiles and social network data. The following chapters then demonstrate the use of geospatial technologies for managing virtual communities, and for monitoring, analyzing and mapping environmental indicators. Finally, the last four chapters address service-oriented architectures, and describe how distributed Web services facilitate the integration of knowledge repositories with geospatial platforms and third-party applications.

I congratulate the authors for their excellent and timely work. The book is not only a comprehensive, interdisciplinary collection of current research; it also introduces visionary concepts and outlines promising avenues for future research.

Patrick J. Hogan
Program Manager, NASA World Wind
worldwind.arc.nasa.gov

Preface

Contrary to early predictions that the Internet will obsolete geography, the discipline is increasingly gaining importance. In a 1998 speech at the California Science Center, former U.S. Vice President Al Gore called for replacing the prevalent desktop metaphor with a "multi-resolution, three-dimensional representation of the planet, into which we can embed vast quantities of geo-referenced data" (Gore 1998). After the successful introduction of three-dimensional geospatial platforms such as NASA World Wind,[1] Google Earth[2] and Microsoft Live Local 3D,[3] achieving the vision of a *Geospatial Web* seems more realistic than ever.

Dubbed the "holy grail of mapping" (Levy 2004), these geobrowsers aggregate and project layers of metadata onto scale-independent spherical globes. They are an ideal platform to integrate (i) cartographic data such as topographic maps and street directories, (ii) geotagged knowledge repositories aggregated from public online sources or corporate intranets, and (iii) environmental indicators such as emission levels, ozone concentrations and biodiversity density. By integrating cartographic data with geotagged knowledge repositories, the Geospatial Web will revolutionize the production, distribution and consumption of media products.

The appearance of geobrowsers in mainstream media coverage (see Chapter 1) increases public acceptance of geospatial technology and improves geospatial literacy, which today exists only among a small portion of highly educated people (Erle et al. 2005). Geospatial literacy includes the ability to understand, create and use geospatial representations for Web navigation, narrative descriptions, problem-solving and artistic expression (Liebhold 2004). In light of the explosive growth and diminished lifespan of information, geospatial literacy is becoming increasingly important, as the thought that needs to be followed in information discovery tasks is often spatial in nature (McCurley 2001). Geobrowsing platforms support such information discovery tasks by allowing users to switch between or integrate a large number of heterogeneous information services.

The 25 chapters contained in this edited volume summarize the latest research on the Geospatial Web's technical foundations, describe information services and collaborative tools built on top of geobrowsers and investigate the environmental, social and economic impacts of knowledge-intensive applications. Supplemental material including author biographies and bibliographic resources is available from the book's official Web site at

www.geospatialweb.com

The book emphasizes the role of contextual knowledge in shaping the emerging network society. Several chapters focus on the integration of geospatial and semantic technology to extract geospatial context from unstructured textual resources; e.g., to automatically identify and map the most relevant content for customized news services. Hybrid models combine such automated services with the advantages of individual and collaborative content production environments – for example by integrating "edited" material from newspapers and traditional encyclopedias with "evolving" content from collaborative Wiki applications.

Automatically annotating content acquired from these different sources creates knowledge repositories spanning multiple dimensions (space, time, semantics, etc.). Geospatial exploration systems will improve the accessibility and transparency of such complex repositories.

Keen competition between software and media companies surrounds the provision of geospatial exploration systems. The platforms are evolving quickly, gaining new functionality, data sources and interface options in rapid succession. But the currently available applications only hint at the true potential of geospatial technology. The Geospatial Web will have a profound impact on managing individual and organizational knowledge. It will not only reveal the context and geographic distribution of a broad range of information services and location-based resources but also help create and maintain virtual communities by matching people of similar interests, browsing behavior or geographic location.

Acknowledgements

This book would not have been possible without the help and contributions of many colleagues. Our first word of appreciation goes to the authors for their excellent work and active participation in the peer-review process. Each chapter was evaluated by three or four referees and revised at least once on the basis of their comments and criticism.

We would like to thank the following colleagues who generously provided additional reviews and feedback: Albert Weichselbraun, Andrea Polli, Andreas Juffinger, Benno Stein, Deana Pennington, Eva Micietova, Fridolin Wild, Herwig Rollett, Joachim P. Hasebrook, Jörg Westbomke, Klaus Leopold, Kostas Karatzas, Marc Van Liedekerke, Markus Strohmaier, Mathias Lux, Michael Granitzer, Nguyen Xuan Thinh, Panos Panagos, S.K. Ghosh, Soenke Dohrn, Stefan Kollarits, Tomas Pitner, Wei Liu, Wolf-Fritz Riekert, Wolfgang Kienreich and Yiwei Cao.

Gabriele Zorn-Pauli is to be commended for her valuable assistance in the editorial process. We would also like to thank the series editors of *Advanced Information and Knowledge Processing*, Xindong Wu and Lakhmi Jain, as well as the staff at Springer for their support and help in the materialization of this book.

Arno Scharl
Klaus Tochtermann Graz, April 2007

Know-Center Graz – Austria's Competence Center for Knowledge Management
Graz University of Technology, Knowledge Management Institute
Inffeldgasse 21a, A-8010 Graz, Austria

www.know-center.at ▪ kmi.tugraz.at ▪ www.idiom.at ▪ www.ecoresearch.net

The Know-Center is funded by the Austrian Competence Center program Kplus under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology (BMVIT), and by the State of Styria. The IDIOM (Information Diffusion across Interactive Online Media) research project is funded by BMVIT and the Austrian Research Promotion Agency (FFG) within the strategic objective FIT-IT Semantic Systems (www.fit-it.at).

Table of Contents

Foreword v
Preface vii
Table of Contents ix
List of Authors xi

FOUNDATIONS OF THE GEOSPATIAL WEB

1 Towards the Geospatial Web: Media Platforms for Managing Geotagged Knowledge Repositories 3
2 Infrastructure for the Geospatial Web 15
3 Imaging on the Geospatial Web Using JPEG 2000 27
4 What's So Special about Spatial? 39

NAVIGATING THE GEOSPATIAL WEB

5 Conceptual Search: Incorporating Geospatial Data into Semantic Queries 47
6 Location-based Web Search 55
7 Ubiquitous Browsing of the World 67
8 Spatiotemporal-Thematic Data Processing for the Semantic Web 79

BUILDING THE GEOSPATIAL WEB

9 A Semantic Approach for Geospatial Information Extraction from Unstructured Documents 93
10 Enhancing RSS Feeds with Extracted Geospatial Information for Further Processing and Visualization 105
11 A Supervised Machine Learning Approach to Toponym Disambiguation 117

GEOSPATIAL COMMUNITIES

12 Geospatial Information Integration for Science Activity Planning at the Mars Desert Research Station 131
13 Inferences of Social and Spatial Communities over the World Wide Web 141
14 Participating in the Geospatial Web: Collaborative Mapping, Social Networks and Participatory GIS 153
15 Sharing, Discovering and Browsing Geotagged Pictures on the World Wide Web 159
16 Supporting Geo-Semantic Web Communities with the DBin Platform: Use Cases and Perspectives 171

ENVIRONMENTAL APPLICATIONS

17 A Geospatial Web Platform for Natural Hazard Exposure Assessment in the Insurance Sector 179
18 Development, Implementation and Application of the WebGIS MossMet 191

19 European Air Quality Mapping through Interpolation with Application to Exposure and Impact Assessment 201
20 Introduction to Ubiquitous Cartography and Dynamic Geovisualization with Implications for Disaster and Crisis Management 209
21 Fire Alerts for the Geospatial Web 215

GEOSPATIAL WEB SERVICES

22 Geospatial Web Services: The Evolution of Geospatial Data Infrastructure 223
23 SWING – A Semantic Framework for Geospatial Services 229
24 Similarity-based Retrieval for Geospatial Semantic Web Services Specified Using the Web Service Modeling Language (WSML-Core) 235
25 Geospatial Data Integration with Semantic Web Services: The eMerges Approach 247

Bibliography 257
Online Resources 287
Index 291

List of Authors

Dipl.-Umweltwiss.
Christian Aden
PhD Candidate
University of Vechta, Chair of Landscape Ecology; Vechta, Germany

Dr
Pragya Agarwal
Lecturer
Department of Geomatic Engineering
University College London
London, UK

Dipl. Inf.
Dirk Ahlers
Research Assistant
OFFIS Institute for Information Technology; Oldenburg, Germany

MAMS
Boanerges Aleman-Meza
Research Assistant
University of Georgia, Computer Science Department, LSDIS Lab; Athens, GA, USA

Dr
Steve Battle
Research Engineer
Hewlett-Packard Laboratories
Bristol, UK

Mr
Torsten Becker
Publicist and Author
Founder of ExploreOurPla.net
Cologne, Germany

Dr
Roderic Bera
Lecturer
Department of Geomatic Engineering
University College London
London, UK

MA
Susan J. Bergeron
PhD Candidate
West Virginia University
Department of Geology and Geography
Morgantown, WV, USA

MD PhD
Daniel C. Berrios
Scientist
University of California, Santa Cruz
NASA Ames Research Center
Moffett Field, USA

Dr
Susanne Boll
Professor of Media Informatics and Multimedia Systems
University of Oldenburg
Oldenburg, Germany

MS
Gabriella Castelli
PhD Candidate
University of Modena and Reggio Emilia, Department of Science and Methodologies of Engineering; Reggio Emilia, Italy

Dr
Steve Cayzer
Research Engineer
Hewlett-Packard Laboratories
Bristol, UK

MSc
Jérôme Chemitte
PhD Candidate
Mission Risques Naturels,
Ecole des Mines de Paris, Pôle Cindyniques
Sophia Antipolis, France

Dr
Christophe Claramunt
Professor and Director
Naval Academy Research Institute
Brest, France

BA MSc
Rob Davies
Partner
MDR Partners
London, UK

BS
Mike Dean
Principal Engineer
BBN Technologies
Arlington, Virginia, USA

Dr
Bruce Denby
Senior Researcher
Norwegian Institute for Air Research
Kjeller, Norway

Dr
Catherine Dolbear
Senior Research Scientist
Ordnance Survey Research Labs
Southampton, United Kingdom

Dr
John Domingue
Deputy Director
Knowledge Media Institute (KMI)
The Open University
Milton Keynes, UK

Mr
Jim Farley
Vice President
Galdos Systems
Vancouver, Canada

Mr
Matthew Fleagle
Senior Technical Writer
LizardTech
Seattle, WA, USA

Dr
Mauro Gaio
Professor of Computer Science
LIUPPA Laboratory
Laboratoire d'Informatique de l'Université de Pau et des Pays de l'Adour; Pau, France

Dr
Linlin Ge
Senior Lecturer
University of New South Wales
School of Surveying and Spatial Information Systems
Sydney, Australia

MS
Michael P. Gerlek
Engineering Manager/Software Architect
LizardTech
Seattle, WA, USA

Dr
Alessio Gugliotta
Research Fellow
Knowledge Media Institute (KMI)
The Open University
Milton Keynes, UK

MSc
Leticia Gutierrez
Ontology Engineer
Essex County Council
Chelmsford, Essex, UK

PhD
Farshad Hakimpour
Research Associate
University of Georgia, Computer Science Department, LSDIS Lab
Athens, GA, USA

Dr
Trevor M. Harris
Eberly Distinguished Professor of Geography
West Virginia University, Department of Geology and Geography
Morgantown, WV, USA

BSc (Hons), BA (Hons)
Glen Hart
Principal Research Scientist and Research Manager
Ordnance Survey Research Labs
Southampton, UK

Mr
Marcel Holy
Student Academic Staff
University of Vechta
Chair of Landscape Ecology
Vechta, Germany

Mgr
Jan Horalek
Air Quality Researcher
Czech Hydrometeorological Institute
Prague, Czech Republic

Dr
Jirí Hrebícek
Professor of Information Systems
Masaryk University
Institute of Biostatistics and Analyzes
Brno, Czech Republic

Mr
You-Heng Hu
PhD Candidate
University of New South Wales
School of Surveying and Spatial Information Systems
Sydney, Australia

MSc
Julien Iris
PhD Candidate
Ecole des Mines de Paris
Pôle Cindyniques
Sophia Antipolis, France

Dipl.-Lök
Krzysztof Janowicz
Research Associate/PhD Student
Münster Semantic Interoperability Lab
Institute for Geoinformatics
University of Münster, Germany

BS
William Kammersell
Software Engineer
BBN Technologies
Arlington, Virginia, USA

PhD
Richard M. Keller
Computer Scientist
NASA Ames Research Center
Intelligent Systems Division
Moffett Field, Mountain View, CA, USA

Dipl.-Umweltwiss.
Lukas Kleppin,
PhD Candidate
University of Vechta
Chair of Landscape Ecology
Vechta, Germany

Dipl.-Landsch.-Ökol.
Eva Klien
Research Associate
Institute for Geoinformatics
University of Münster
Münster, Germany

Dr
Milan Konecný
Associate Professor and President of the International Cartography Association
Masaryk University, Laboratory of Cartography & Geography
Brno, Czech Republic

BA MA CGS (GIS/LIS)
Athanasios Tom Kralidis
Senior Systems Scientist
Environment Canada
Toronto, Ontario
Canada

Mr
Ron Lake
CEO
Galdos Systems
Vancouver, British Columbia
Canada

PhD
Julien Lesbegueries
PIV Project Member
LIUPPA Laboratory
Laboratoire d'Informatique de l'Université de Pau et des Pays de l'Adour; Pau, France

PhD
Pierre Loustau
PIV Project Member
LIUPPA Laboratory
Laboratoire d'Informatique de l'Université de Pau et des Pays de l'Adour; Pau, France

PhD
Marco Mamei
Researcher
University of Modena and Reggio Emilia
Department of Science & Methodologies of Engineering; Reggio Emilia, Italy

Dr
Ernesto Marcheggiani
Post Doc Researcher
Università Politecnica delle Marche, Department of Applied Science of Complex Systems, DiSASC; Università Politecnica delle Marche; Ancona, Italy

MA
Graeme McFerren
Senior Researcher
Information & Communications Technology for Earth Observation Group
Meraka Institute, CSIR
Pretoria, South Africa

Dr
Christian Morbidoni
Post Doc Researcher
Semantic Web and Multimedia Group (SEMEDIA); D.E.I.T, Università Politecnica delle Marche; Ancona, Italy

Dr
Aldo Napoli
Researcher
Ecole des Mines de Paris, Pôle Cindyniques
Sophia Antipolis, France

Ing.
Michele Nucci
PhD Candidate
Semantic Web and Multimedia Group (SEMEDIA); D.E.I.T, Università Politecnica delle Marche; Ancona, Italy

MSc
Matthew Perry
Research Assistant
University of Georgia, Computer Science Department, LSDIS Lab
Athens, GA, USA

Dr
Roland Pesch
Research Assistant
University of Vechta, Chair of Landscape Ecology; Vechta, Germany

BSc
Marc Richardson
Semantic Web Researcher
Next Generation Web Research
BT Group Chief Technology Office
Ipswich, UK

Dipl.-Eng.
Dumitru Roman
Researcher
DERI Innsbruck
University of Innsbruck
Innsbruck, Austria

MSc
Stacey Roos
Researcher
Information & Communications Technology for Earth Observation Group
Meraka Institute, CSIR
Pretoria, South Africa

MS
Alberto Rosi
PhD Candidate
University of Modena and Reggio Emilia
Department of Science and Methodologies of Engineering; Reggio Emilia, Italy

Mr
L. Jesse Rouse
PhD Candidate
West Virginia University, Department of Geology and Geography
Morgantown, WV, USA

BA MSc
Mary Rowlatt
Customer Relations Manager
Essex County Council
Chelmsford, Essex, UK

Dr
Christian Sallaberry
Assistant Professor of Computer Science
LIUPPA Laboratory
Laboratoire d'Informatique de l'Université de Pau et des Pays de l'Adour; Pau, France

DDr
Arno Scharl
Professor of New Media and Knowledge Management
Know-Center and Graz University of Technology, Knowledge Management Institute; Graz, Austria

Dr
Gunther Schmidt
Research Assistant
University of Vechta
Chair of Landscape Ecology
Vechta, Germany

Dr
Winfried Schröder
Professor
University of Vechta
Chair of Landscape Ecology
Vechta, Germany

PhD
Amit Sheth
Professor of Computer Science
University of Georgia, Computer Science Department, LSDIS Lab
Athens, GA, USA

Dr Ing
Maarten Sierhuis
Senior Scientist
Research Institute for Advanced Computer Science, NASA Ames Research Center, Moffett Field
Mountain View, CA, USA

Msc
Peter A. M. de Smet
Senior Policy Researcher
Netherlands Environmental Assessment Agency; Bilthoven, The Netherlands

BSc
Sandra Stinčić
Semantic Web Researcher
Next Generation Web Research
BT Group Chief Technology Office
Ipswich, UK

BA MSc
Vlad Tanasescu
PhD Student
The Open University
Knowledge Media Institute (KMi)
Milton Keynes, UK

MSc
Andrew Terhorst
Research Group Leader
Information and Communications Technology for Earth Observation Group
Meraka Institute, CSIR
Pretoria, South Africa

Dr
Carlo Torniai
Researcher
Multimedia Integration and Communication Center, Universita' di Firenze
Firenze, Italy

Dr
Giovanni Tummarello
Post Doc Researcher
Semantic Web and Multimedia Group (SEMEDIA); D.E.I.T, Università Politecnica delle Marche; Ancona, Italy

Mr
Marc Wick
Software Engineer
Project Lead Geonames.org
St. Gallen, Switzerland

PhD
Franco Zambonelli
Professor
University of Modena and Reggio Emilia
Department of Science and Methodologies of Engineering; Reggio Emilia, Italy

Chapter 1

Towards the Geospatial Web: Media Platforms for Managing Geotagged Knowledge Repositories

Arno Scharl

Abstract. *International media have recognized the visual appeal of geobrowsers such as NASA World Wind and Google Earth, for example, when Web and television coverage on Hurricane Katrina used interactive geospatial projections to illustrate its path and the scale of destruction in August 2005. Yet these early applications only hint at the true potential of geospatial technology to build and maintain virtual communities and to revolutionize the production, distribution and consumption of media products. This chapter investigates this potential by reviewing the literature and discussing the integration of geospatial and semantic reference systems, with an emphasis on extracting geospatial context from unstructured text. A content analysis of news coverage based on a suite of text mining tools (webLyzard) sheds light on the popularity and adoption of geospatial platforms.*

1.1 Introduction

Historically, media technology enters the market via new types of content that drive adoption and validate emerging business models. For true media innovation to have human impact, however, it must affect the imagination – creating an associated magic "behind the eyeballs" that changes people's behavior in their commercial, academic and personal environments (Stapleton and Hughes 2006). The following hypothetical scenario outlines how geospatial technology may radically change working environments, impact workflows within and across organizations, and enrich the interaction between content providers and their target audience.

Kathryn O'Reilly is a freelance editor who sells her ability to gather, filter and prioritize electronic content. In a virtual world built on contextualized information spaces, Kathryn seamlessly switches between geographic and semantic topologies. She begins her typical working day floating in the virtual space above Earth, ready to navigate the globe and semantic structures via subtle movements of her eyes. An extensive portfolio of add-on functionality is accessible through haptic devices. From her elevated position, Kathryn not only observes the rise and decay of topics, but also the unfolding of social structures based on the unique social networks of her friends and business contacts. Across these networks she builds and shares her knowledge repository and composes media products that are continuously being validated and enriched by the latest news feeds and third-party sources.

The underlying content management system tailors the format of her articles to the changing preferences of her regular readers. Kathryn adds, selects, categorizes, aggregates, filters and extrapolates information along multiple dimensions, with minimal cognitive requirements. She can structure her daily workflows, access archives of historic

textual and multimedia data and customize her virtual environment. Adaptive communication services allow her to interact with predefined or dynamically assembled groups of like-minded individuals. At any point in time, Kathryn may use portions of the information space to initialize simple what-if scenarios or advanced socioeconomic simulations, investigating the complex interplay among computer-generated avatars, automated information services and real-world participants.

In the words of McLuhan, media as an extension of ourselves provide new transforming vision and awareness (McLuhan 1964). In the early 1940s, the first images of Earth from space eroded limitations to human perception, triggered profound self-reflexive experiences (DeVarco 2004) and revitalized public desire to preserve a beautiful but vulnerable planet (Biever 2005). Thanks to human space exploration, therefore, most users will instantly recognize our planet and find it an intuitive and effective metaphor to access and manage geotagged information: "There it is, that good old pale blue dot in all its earthly glory, right there on your computer screen. It's a familiar sight, even from a sky-high perspective experienced only by astronauts and angels" (Levy 2004, 56).

As the concepts of "desktop," "village" and "landscape" have shown, well-known interface metaphors are powerful instruments to gain market acceptance (Fidler 1997). Geobrowsers promote the "planet" metaphor by providing users with an accurate visual representation and allowing them to browse geospatial data from a satellite perspective. Using standardized services such as the bitmap-based Web Map Service (WMS)[4] and the vector-based Web Feature Service (WFS),[5] image tiles and vector data including geo-positioning information are retrieved from a central server, arranged into real-time mosaics and mapped onto three-dimensional representations of the globe. Altering the field-of-view angle allows users to switch between detailed views and highly aggregated representations. Users can effortlessly zoom from Blue Marble Data[6] at a 1-kilometer-per-pixel rate, for example, to the detailed mosaic of LandSat 7 Data[7] at 15 meters per pixel (Hogan and Kim 2004). Adding the option to tilt the display relative to the spectator's point of view adds altitude as a third dimension.

Given the potential of the "planet" metaphor, academia and industry alike call for a new generation of geospatial interfaces with simple yet powerful navigational aids that facilitate the real-time access and manipulation of geospatially and semantically referenced information.

1.2 Geospatial Reference System

Observing, aggregating and visualizing human behavior are common activities, from tracking customers in retailing outlets to monitoring traffic in congested urban areas, or analyzing the clickstreams of online shoppers based on Web server logfiles (Scharl 2001). Prior to the advent of the Global Positioning System (GPS) and Radio Frequency Identification (RFID), the lack of appropriate technology to pinpoint a user's precise location restricted the functionality of many applications. Nowadays, aggregated visualizations of individual actions are a familiar sight, as geobrowsers redefine the look and feel of user interfaces and leverage the knowledge about a user's precise location to unlock organized indices to the physical world (Kendall 2005).

Information retrieval research has also discovered geobrowsers as an effective platform to identify and access relevant information more effectively. An increasing number of applications use geospatial extensions for specifying queries and structuring the presentation of results.

Most providers of geobrowsing platforms offer *Application Programming Interfaces (APIs)* or *XML scripting* to facilitate building third-party online services on top of their platforms (Roush 2005). Multiple layers of icons, paths and images can be projected via these services. Such visual elements are scaled, positioned on the globe, and linked to (Web) documents,[8] photo collections[9] and other external resources (Neches et al. 2001). Latitude and longitude variables determine the symbols' position, while distance above surface values specify whether symbols hover above ground. A good example is data from NASA's Moderate Resolution Imaging Spectroradiometer (MODIS),[10] providing daily updated planetary imagery at resolutions up to 250 meters per pixel that documents natural events such as fires, floods, storms and volcanic activity (Hogan and Kim 2004). The left screenshot of Figure 1.1 shows an MODIS overlay of Hurricane Katrina as of August 29, 2005.

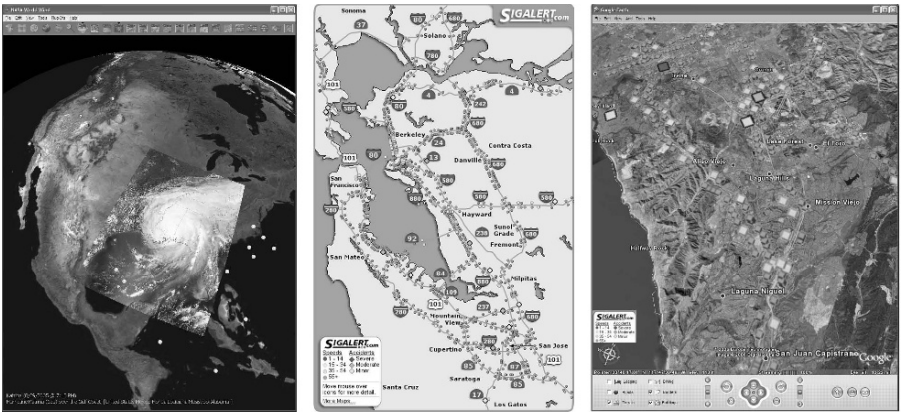


Figure 1.1: Data integration with NASA World Wind and Google Earth

The availability of APIs is largely responsible for the growing popularity of mash-ups. Often released by individuals or the open-source community, mash-ups combine publicly available data and interface services from multiple providers into an integrated user experience (Hof 2005). The map in the center of Figure 1.1 displays the original Sigalert.com service that aggregates real-time traffic data from the San Francisco Bay Area. The screenshot on the right exemplifies the idea of a mash-up, using the Sigalert.com traffic data to project symbols for accidents and current traffic speeds onto the Google Earth representation of Southern Los Angeles.[11]

1.2.1 Extraction and Disambiguation of Geospatial Context

Concentrated efforts are under way to geotag as much existing information as possible. Geotagging refers to the process of assigning geospatial context information, ranging from specific point locations to arbitrarily shaped regions. Different sources of geospatial context information for annotating Web resources often complement each other in real-world applications (McCurley 2001):

- Annotation by the author, manually (Daviel and Kaegi 2003) or through location-aware devices such as car navigation systems, RFID-tagged products and GPS-enabled cellular handsets. These devices geotag information automatically when it is being created.

- Determining the location of the server – e.g., by querying the Whois[12] database for domain registrations, monitoring how Internet traffic is routed on a macro level, or by analyzing the domain of a Web site for additional cues.
- Automated annotation of existing documents. The processes of recognizing geographic context and assigning spatial coordinates are commonly referred to as *geoparsing* and *geocoding*, respectively.

Once geospatial context information becomes widely available, any point in space will be linked to a universe of commentary on its environmental, historical and cultural context, to related community events and activities and to personal stories and preferences. Even locative spam will become a common phenomenon (Erle et al. 2005) with the widespread introduction of location-based services, geospatial gaming environments and other commercial applications.

At present, however, many metadata initiatives still suffer from the chicken and egg problem, wishing that existing content was retrofitted with metadata (McCurley 2001). This "capture bottleneck" results from the beneficiaries' lack of motivation to devote the necessary resources for providing a critical mass of metadata (Motta et al. 2000). Geotagging projects are no exception. Acknowledging calls to automate the semantic annotation of documents (Benjamins et al. 2004; Domingue and Motta 2000), the following sections focus on the third category, the automated geoparsing and geocoding of existing Web resources – online news, for example, or other types of unstructured textual data found on the Web.

1.2.1.1 Geoparsing

All human artifacts have a location history, which commonly includes a creation location and a current location (Spohrer 1999). Given the availability of metadata, geospatial applications can map the whole life cycle of such artifacts. Electronic resources contain metadata as explicit or implicit geographic references. This includes references to physical features of the Earth's surface such as forests, lakes, rivers and mountains, and references to objects of the human-made environment such as cities, countries, roads and buildings (Jones et al. 2001). Addresses, postal codes, telephone numbers and descriptions of landmarks also allow us to pinpoint exact locations (Ding et al. 2000; McCurley 2001).

At least 20 percent of Web pages contain easily recognizable and unambiguous geographic identifiers (Delboni et al. 2005). News articles are particularly rich in such identifiers, since they usually discuss the location where an event took place, or where it was reported from (Morimoto et al. 2003). The BBC article "Vienna Marking Mozart Milestone" (Bell 2006), for example, has a target geography of EUROPE/AUSTRIA/VIENNA and a source geography of EUROPE/UNITED KINGDOM/LONDON. In addition to target and source geography (Amitay et al. 2004), natural language processing can also be used to extract the geographic scope (i.e., intended reach) of Web resources (Wang et al. 2005).

Identifying and ranking spatial references by semantically analyzing textual data is a subset of the more general problem of *named entity recognition*, which locates and interprets phrasal units such as the names of people, organizations and places (Cowie and Lehnert 1996; Weiss et al. 2005). As with most named entity recognition tasks, false positives are inevitable – e.g., documents that quote addresses unrelated to the their actual content (Morimoto et al. 2003).

Ambiguity, synonymy and changes in terminology over time further complicate the geoparsing of Web documents (Amitay et al. 2004; Kienreich et al. 2006; Larson 1996). Identical lexical forms refer to distinct places with the same name (VIENNA refers to the capital of Austria as well as a town in Northern Virginia, USA) or have geographic and non-geographic meanings: TURKEY (large gallinaceous bird; bi-continental country between Asia and Europe), MOBILE (capable of moving; city in Alabama, USA), or READING (processing written linguistic messages; town in Massachusetts, USA). Geoparsers also need to correctly process references to identical or similar places that may be known under different names, or may belong to different levels of administrative or topographical hierarchies (Jones et al. 2001).

1.2.1.2 Geocoding

Once a location has been identified, precise spatial coordinates – latitude, longitude and altitude – can be assigned to the documents by querying structured geographic indices (gazetteers) for matching entries (Hill et al. 1999; Tochtermann et al. 1997). This process of associating documents with formal models is also referred to as document enrichment (Domingue and Motta 2000; Motta et al. 2000). Examples of formal geographic models are the Geographic Names Information System (GNIS),[13] the World Gazetteer,[14] the classifications of the United Nations Group of Experts on Geographical Names,[15] the Getty Thesaurus of Geographic Names[16] and ISO 3166-1 Country Codes.[17]

While simple gazetteer lookup has the advantage of being language-independent, advanced algorithms consider lexical and structural linguistic clues as well as contextual knowledge contained in the documents; e.g., dealing with ambiguity by removing stop-words, identifying references to people and organizations (Clough 2005) and applying contextual rules like "single sense per document" and "co-occurring place names indicate nearby locations". Each identified reference is assigned a probability *P(name, place)* that it refers to a particular place (Amitay et al. 2004). The location that receives the highest probability is then assigned a canonical taxonomy node such as EUROPE/AUSTRIA/VIENNA; 48°14' N, 16°20' E.

1.2.2 Managing Geospatial Context

Standardized metadata frameworks often include geospatial attributes like the Dublin Core Metadata Initiative's "Coverage" tag (McCurley 2001).[18] The need for controlled vocabularies and shared meaning suggests that ontologies are going to play a key role in managing geospatial context information. While conflicting definitions of "ontology" abound (Guarino 1997), most researchers agree that the term refers to a designed artifact formally representing shared conceptualizations within a specific domain (Gahleitner et al. 2005; Jarrar and Meersman 2002).

Geo-ontologies encode geographical terms and semantic relationships such as containment, overlap and adjacency (Tochtermann et al. 1997). Spatially aware search engines use ontological knowledge for query term expansion and disambiguation, relevance ranking and Web resource annotation (Abdelmoty et al. 2005). Geo-ontologies can either be represented through generic markup languages like the Web Ontology Language (OWL)[88] endorsed by the World Wide Web Consortium (Horrocks et al. 2003; Smith et al. 2004) or more specific approaches like the Geography Markup Language (GML)[28] developed by the Open Geospatial Consortium (Lake et al. 2004; see Chapter 2 "Infrastructure for the Geospatial Web" for a more detailed discussion).

1.3 Semantic Reference System

Geospatially referenced information enables geobrowsers to map annotated content units from various sources, track human activities and visualize the structure and dynamics of virtual communities. But geobrowsers can also serve as a generic image rendering engine to project other types of imagery. Diverting them from their traditional purpose and connecting them to *semantically* referenced information, they can be used to visualize *knowledge planets* based on layered thematic maps. Such maps are visual representations of semantic information spaces based on a landscape metaphor (Chalmers 1993).

Generally, two sets of information need to be integrated and mapped to latitude and longitude – image tiles and terrain information. Knowledge planets are generated by orthographically projecting and tiling thematic maps. The planet metaphor allows visualizing massive amounts of textual data. At the time of map generation, the knowledge planet's topology is determined by the content of the knowledge base. The peaks of the virtual landscape indicate abundant coverage on a particular topic, whereas valleys represent sparsely populated parts of the information space.

Extending the planet metaphor, search results can be visualized as cities, landmarks or other objects of the manmade environment. Zooming provides an intuitive way of selecting the desired level of aggregation. Unique resource identifiers link concepts embedded in the thematic maps to related news articles, encyclopedia entries or papers in scientific journals. With such a query interface that hides the underlying complexity, exploring complex data along multiple dimensions is as intuitive as using a geobrowser to get a glimpse of the next holiday destination.

VisIslands, a thematic mapping algorithm similar to SPIRE's Themescape (Wise 1999) and its commercial successor Cartia/Aureka (see Figure 1.2),[19] supports dynamic document clustering (Andrews et al. 2001; Sabol et al. 2002). Initially, the document set is pre-clustered using hierarchical agglomerative clustering (Jain et al. 1999), randomly distributing the cluster centroids in the viewing rectangle. The documents belonging to each cluster, as determined by the pre-clustering, are then placed in circles around each centroid. The arrangement is fine-tuned using a linear iteration force-directed placement algorithm adapted from Chalmers (1996). The result resembles a contour map of islands. Fortunately, algorithms based on force models easily generalize to the knowledge planets' spherical geometries.

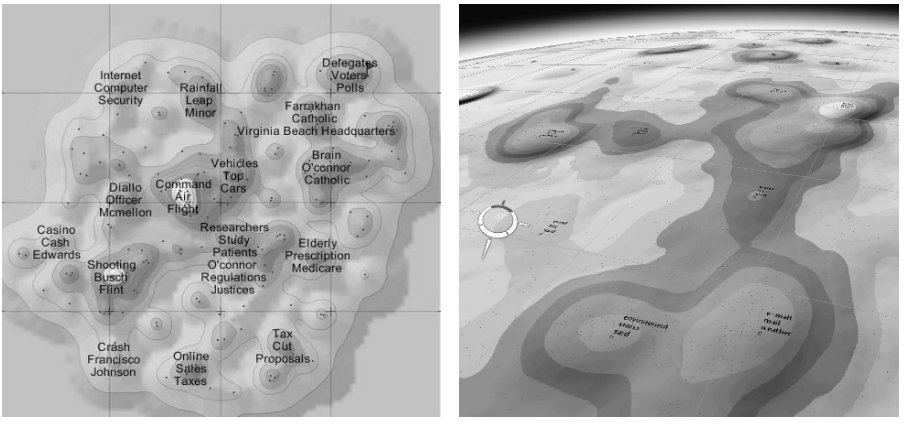


Figure 1.2: Thematic mapping exemplified by Cartia (left) and a knowledge planet prototype (right)

The IDIOM (Information Diffusion across Interactive Online Media)[20] research project extends and refines the VisIslands thematic mapping component to improve throughput and scalability, generate layered thematic maps, and provide a Web Map Service (WMS) that serves these maps as image tiles for various geobrowsing platforms. The NASA World Wind screenshot of Figure 1.2 shows an early prototype of this service. The transition from two-dimensional thematic maps to three-dimensional knowledge planets poses conceptual and technical challenges – the initial arrangement of major concepts, for example, which should be guided by domain ontologies. Users will expect a consistent experience when rotating the planet. This requires a seamless flow of concepts when crossing the planet's 0° meridian line. The same principle applies to zooming operations. Analogous to Landsat-7 data, multiple layers of thematic maps in different resolutions and with appropriate sets of captions have to be synchronized with each other.

1.4 Geospatial Publishing

Technological convergence and the move towards digital media continue to drive today's newsrooms (Pavlik 1998). While many innovations that gain ground in the media industry are largely invisible to the end user, geobrowsers directly impact the consumption of news media, change mainstream storytelling conventions and provide new ways of selecting and filtering news stories. By facilitating the access of annotated knowledge repositories, geobrowsers set the stage for the Geosptial Web as a new platform for content production and distribution.

1.4.1 Content Production and Distribution

Hybrid models of individual and collaborative content production are particularly suited for geobrowsers, which can integrate and map *individual sources* (monographs, commentaries, blogs), *edited sources* (encyclopedias, conference proceedings, traditional news services), *evolutionary sources* (Wiki applications, open-source project documentations) and *automated sources* (document summarizers, news aggregators). Geobrowsing technology not only impacts the production of content, but also its distribution, packaging and consumption. When specifying preferences for personalized news services, for example, geobrowsers are effective tools to pinpoint locations and specify geographic areas to be covered by the news service.

Personalized news services require content that is correctly annotated along at least three dimensions: (i) *spatial* – e.g., distinguishing between source and target geography; see Section 1.2.1; (ii) *semantic* – e.g., assigning the most relevant concepts from a controlled vocabulary; and (iii) *temporal* – e.g., adding timestamps for the reported event, the initial publication and subsequent revisions. Online news can be organized, indexed, searched and navigated along these dimensions:

- The *geographical scope* of an article allows filtering and prioritizing electronic content in line with the user's area of interest, which is often different from his or her actual location.
- *Topical similarity* is another common dimension to tag and filter news content, often matched against user-specific degree of interest functions.
- Finally, by adding a *temporal dimension* through time distribution graphs or visual animation, change over time along any other dimension can be captured; e.g., the unfolding of events, news distribution patterns or the inter-individual propagation of personal messages.

Most geographic information systems, however, treat time as an attribute rather than a separate dimension (Johnson 2004). This is about to change as geobrowsing platforms prepare the transition towards a fully functional Geospatial Web. Dynamic queries, interactive time displays and playback controls will enable users to identify the rise and decay of topics – the diffusion of news coverage on natural disasters, for example, or the impact of political events.

The simplest way of developing a news browser is to combine existing data sources and interface services (see Section 1.2). The news summary[21] on the left side of Figure 1.3 receives the News Feeds of Associated Press,[22] processes this stream of data with the Yahoo! Geocoding API[23] and displays the results via the Google Maps interface. More specific requirements or research interests often result in standalone applications. The second screenshot of Figure 1.3 (Rüger 2005) shows a geotemporal news browser that allows users to search a news database via query terms and time-interval sliders and presents matching articles mapped onto a region of interest. It follows Shneiderman and Plaisant's (2004) information seeking mantra: generate an overview, provide zooming and filtering, and present details on demand. These guidelines avoid clutter in the display, which results from projecting too many content items from a large knowledge repository simultaneously (Larson 1996). Instead of showing the complete set of available news items, for example, a user may wish to restrict the display to articles on climate change that were published in the online editions of Italian newspapers within the last 48 hours.

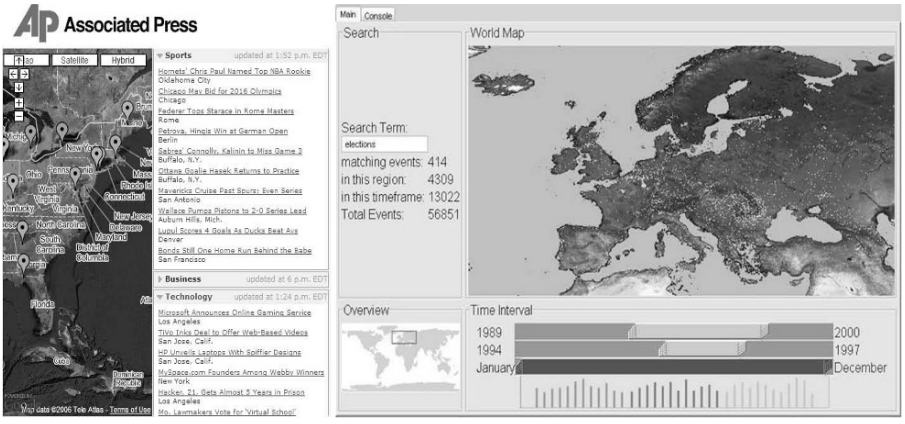


Figure 1.3: Interfaces for accessing geo-referenced news archives

1.4.2 Economic Implications

In light of the observable technical trends, irrespective of their reach and target audience, newsrooms will have to come to terms with metadata (Schutzberg 2005). The widespread availability of metadata will drive the transition towards the Geospatial Web. Emerging geospatial technology supports restructuring processes within the media sector, enhances the workflows of virtual newsrooms and promotes locally dispersed content production. It also facilitates the distribution of (customized) electronic content, which is usually characterized by network effects. Metcalfe's law describes such effects by stipulating that the aggregate value of networks increases with approximately the square number of adopters (Swann 2002), which suggests first-mover advantages and lock-in effects due to high switching

costs once a network technology dominates the market. Consequently, successful business strategies for providers of geobrowsing platforms and distributors of media products built on top of these platforms use innovation to attract and retain users, quickly grow a community of like-minded individuals around a new technology and successively enlarge this community through synergy effects with other products and services (Wilk 2005). The rise of Google Earth as outlined in Section 1.4.3 and the ever-increasing number of mash-ups leveraging this platform exemplify a successful implementation of this strategy.

But the race to provide the dominant geospatial platform is far from over and might trigger a new standard war (Google's purchase of Keyhole[24] and Microsoft's purchase of GeoTango[25] and Vexcel[26] demonstrate the perceived strategic potential of three-dimensional platforms for aerial imagery). Strong network effects in markets with powerful positive feedback loops tend to increase the likelihood and intensity of standard wars. In addition to the first-mover advantage of controlling a large base of loyal or locked-in customers, success factors in a standard war include brand name and reputation, intellectual property rights, the ability to innovate, manufacturing capabilities and strength in complements (Shapiro and Varian 1999).

For the Geospatial Web, such complements range from repositories of geotagged documents and user-generated content (e.g., tags and other types of annotation) to location-based services and third-party applications (e.g., simulation games within a geospatial context). With its Flight Simulator,[27] for example, Microsoft looks back on more than 25 years of developing a successful geospatial game engine. Considering its unique position in the operating systems market and large base of locked-in customers, it does not come as a surprise that the company has joined the race to provide the underlying infrastructure for a three-dimensional Geospatial Web. This strategy has worked before. From the first browser war fought against Netscape in the 1990s, Microsoft is known for its "embrace and extend" strategy – imitating technological advances and successfully incorporating them into its flagship products (Shapiro and Varian 1999). It remains to be seen whether the three-dimensional capabilities and high-resolution city textures of Microsoft Virtual Earth 3D will suffice in light of Google's dominance in the search engine market and obvious opportunities to geo-enable the popular and rapidly growing portfolio of Google services.

While platform providers hope to become the substratum upon which all types of electronic content are layered (Levy 2004), first-mover advantages gained through network effects might allow innovative media companies to dominate the information spaces built on top of these platforms. The content management systems of media companies often contain rich geospatial annotations, reflecting both the source and target geography of articles. For articles without geospatial references or only partial annotations, geotagging as outlined in Section 1.2.1 can add the missing information.

Previous geotagging research has developed methods not only to identify a location referenced in a Web resource but also to capture the geographical distribution of its target audience. The *geographical scope* describes the geographical area that its creator intends to reach (Ding et al. 2000). Distinguishing globally relevant material from publications targeting the national, state or city level is particularly relevant, as virtually all media planning models consider gross impressions, reach and frequency of media products (Cannon 2001).

1.4.3 Media Coverage on Geospatial Platforms

Geo-informatics represents an established discipline that has created an industry with remarkable revenues (Wilk 2005), largely hidden from the public eye. The launch of powerful yet intuitive-to-use geobrowsers has increased public awareness of geospatial technology considerably. Spurred by space photography, global satellite positioning, mobile phones, adaptive search engines and new ways of annotating Web content, the "ancient art of cartography is now on the cutting edge" (Levy 2004, 56). Many current articles shine a spotlight on geospatial technologies, describe trends in mobile geospatial applications, investigate the emerging industry of local search or report unusual objects found on satellite images.

In the past, the process of collecting and analyzing such articles was time-consuming and expensive and often yielded incomplete information. Nowadays, information is readily available online, allowing for inexpensive, fast and topical research. As traditional media extend their dominant position to the online world, analyzing their Web sites reflects an important portion of Web content that the average Internet user accesses. On a macro level, analysts gain insights into publicity through incidental news coverage by monitoring information flows across media sites (Scharl et al. 2005). On a micro level, documents retrieved from Web sites contain valuable information about trends and organizational strategies (Scharl 2000).

To investigate the media coverage on geospatial platforms, 129 Web sites were sampled in quarterly intervals between May 2005 and January 2006, drawing upon the *Newslink.org*, *Kidon.com* and *ABYZNewsLinks.com* directories to compile a list of international media sites from seven English-speaking countries: United States, United Kingdom, Canada, Australia, South Africa, New Zealand and Ireland. The *webLyzard.com* crawler followed the Web sites' hierarchical structure until reaching 50 megabytes of textual data, a limit that helped reduce the dilution of top-level information by content in lower hierarchical levels (Scharl 2004). Updated news articles often result in multiple versions of the same content (Kutz and Herring 2005). The system thus identified and removed redundant segments like headlines and news summaries, whose appearance on multiple pages would otherwise distort frequency counts.

Media attention was calculated as the relative number of references to a platform (in occurrences per million tokens). A pattern matching algorithm processed a list of regular expressions, considering common term inflections while excluding potentially ambiguous terms. Table 1.1 categorizes these regular expressions into references to either three-dimensional (3D) or two-dimensional (2D) platforms.

Table 1.1: Regular expression query for geospatial platforms

Geospatial Platforms (3D)	Geospatial Platforms (2D)
(earth|globe|planet)(-|)?(browser|tool|viewer)s?	map(-|)?(browser|tool|viewer)s?
(microsoft|msn?)(-|)(visual|virtual)(-|)?earth	(microsoft|msn?) map(-|)?point
(virtual|digital)(-|)?(earth|globe|planet)s?	google(-|)?(local|maps?)
geo(fusion|matrix)	map(-|)?quest
google(-|)?earth	map(-|)?machine
keyhole(-|)?(2|earthviewer|pro|inc|markup)	windows live(-|)?local
nasa(-|)?world(-|)?wind	yahoo!?(-|)?maps?
terrafly	parc map viewer
terra(-|)?(suite|explorer|builder|gate)	terrain(-|)?(browser|tool|viewer)s?
skyline software	
world(-|)?wind central	
geo(-|)?tango	

Figure 1.4 summarizes the number of occurrences identified by the pattern-matching algorithm. Between Q2/2005 and Q1/2006, coverage on 2D and 3D platforms increased significantly by more than 300 and 1,100 percent, respectively (Wilcoxon Signed Ranks; $p < 0.05$). In the second quarter of 2005, coverage on 2D platforms exceeded coverage on their 3D counterparts (Mann-Whitney; $p < 0.05$). Results from the first quarter of 2006 showed a different picture. There was no significant difference between the categories, although 3D platforms took a slight lead with an average relative frequency of exactly one occurrence per million tokens. Receiving 83 percent of the coverage, Google Earth has been the primary driver behind the observable increase in popularity. This represents a remarkable feat with a product only launched in June 2005, not receiving any mentions in the second quarter of 2005. As of January 2006, MapQuest still dominated the 2D category with 46 percent of total coverage, but Google Maps was catching up rapidly with a share of 44 percent.

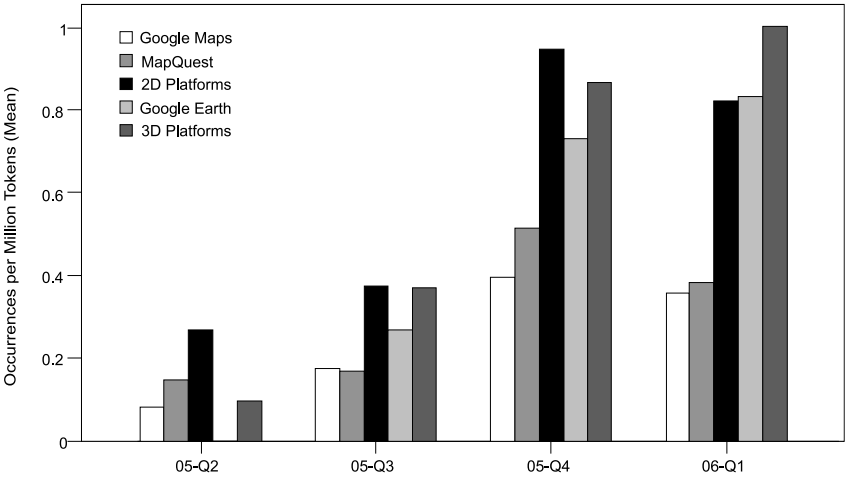


Figure 1.4: Media coverage of geospatial platforms

1.5 Conclusions and Outlook

By integrating cartographic geodata with geotagged hypermedia, the Geospatial Web "may ultimately be the big disruptive innovation of the coming decade" (Erle et al. 2005, xxv). As such, it will serve as a catalyst of social change and enabler of a broad range of as yet unforeseen applications.

The introduction of geobrowsing platforms has popularized the process of "annotating the Planet" (Udell 2005). This chapter outlined the underlying technology, discussed methods to "geo-enable" existing knowledge repositories through parsing geospatial references, and presented several geospatial applications in a media context. A quarterly snapshot of international media coverage revealed the increasing popularity of geospatial technology, particularly as far as three-dimensional platforms are concerned.

Science and technology's accelerated advancement demands constant media innovation, from idea to utility (Stapleton and Hughes 2006). In this competitive environment, geography is emerging as a fundamental principle for structuring the Web (Roush 2005) – a principle that yields the world's knowledge through the lens

of location (Levy 2004, 58). The strategy of adding location metadata to existing databases and accessing the vast amounts of information stored in these databases via geospatial services weds physical and virtual spaces, deepens our experiences of these spaces and incorporates them into our everyday lives (Roush 2005). Coupling tagged knowledge repositories with satellite surveillance and other real-time data sources is a further step towards the *Earth as Universal Desktop*, an idea widely popularized in Neal Stephenson's 1992 novel "Snow Crash":

"A globe about the size of a grapefruit, a perfectly detailed rendition of Planet Earth, hanging in space at arm's length in front of his eyes. … It is a piece of CIC [Central Intelligence Corporation] software called, simply, Earth. It is the user interface that CIC uses to keep track of every bit of spatial information that it owns … It's not just continents and oceans. It looks exactly like the Earth would look from a point in geosynchronous orbit directly above L.A., complete with weather systems – vast spinning galaxies of clouds, hovering just above the surface of the globe, casting gray shadows on the oceans and polar ice caps, fading and fragmenting into the sea. … The computer, bouncing low-powered lasers off his cornea, senses this change in emphasis, and then Hiro gasps as he seems to plunge downward toward the globe, like a space-walking astronaut who has just fallen out of his orbital groove." (Stephenson 1992, 100ff.)

Besides changing individual working environments, geobrowsers are ideally suited for creating and maintaining location-aware communities, bringing people together who share common needs or desires – e.g., communities of friends and social contacts, gaming enthusiasts, political activists or professional acquaintances. Within these communities, geospatial technology helps analyze topics of interest, from the state of the environment to political campaigns, demographic disparity, the progress of civil and urban planning efforts or the structure and efficiency of telecommunications or transportation networks (Erle et al. 2005).

The popularity of contextual advertising and location-based services indicates the technology's remarkable commercial potential. For marketers exploring new media for emerging business opportunities, for instance, the Geospatial Web is "the equivalent of a virgin continent waiting to be planted with billboards" (Roush 2005, 58f.). But established media companies often base strategic decisions on repeating financial successes, a practice that discourages radical innovation (Stapleton and Hughes 2006) and favors nondisruptive technologies. The fact that geospatial technology is compatible with current Internet communication models might help explain its unprecedented rate of adoption, from both organizational and individual perspectives. It integrates well with current protocols and therefore does not replace but complements established modes of navigating Internet resources. This process goes hand-in-hand with the transition towards the Web 2.0, a term that describes advances in Web technology governed by strong network effects and the harnessing of collective intelligence through customer-self service and algorithmic data management (O'Reilly 2005).

Acknowledgements. The author wishes to thank colleagues at the Know-Center and at Vienna University of Economics & Business Administration for their feedback on earlier versions of this chapter. Particular thanks go to Albert Weichselbraun for his help in acquiring the media coverage data (Section 1.4.3). The Know-Center is funded by the Austrian Competence Center program K+ under the auspices of the Austrian Ministry of Transport, Innovation & Technology (BMVIT), and the State of Styria. The IDIOM Project (www.idiom.at) is funded by BMVIT and the Austrian Research Promotion Agency within the strategic objective FIT-IT (www.fit-it.at).

Chapter 2
Infrastructure for the Geospatial Web

Ron Lake • Jim Farley

Abstract. *Geospatial data and geoprocessing techniques are now directly linked to business processes in many areas. Commerce, transportation and logistics, planning, defense, emergency response, health care, asset management and many other domains leverage geospatial information and the ability to model these data to achieve increased efficiencies and to develop better, more comprehensive decisions. However, the ability to deliver geospatial data and the capacity to process geospatial information effectively in these domains are dependent on infrastructure technology that facilitates basic operations such as locating data, publishing data, keeping data current and notifying subscribers and others whose applications and decisions are dependent on this information when changes are made. This chapter introduces the notion of infrastructure technology for the Geospatial Web. Specifically, the Geography Markup Language (GML) and registry technology developed using the ebRIM specification delivered from the OASIS consortium are presented as atomic infrastructure components in a working Geospatial Web.*

2.1 What Is the Geospatial Web?

This article considers the technical foundations for the development, evolution and deployment of a Geospatial Web. For the purposes of this discussion, the Geospatial Web is an integrated, discoverable collection of geographically related Web services and data that spans multiple jurisdictions and geographic regions. In a broad sense, the Geospatial Web refers to the global collection of general services and data that support the use of geographic data in a range of domain applications. Regional and/or domain-specific expressions of the global Geospatial Web exist as well. The global, national, state/provincial or local Spatial Data Infrastructure, or SDI (NRC 1993), are each instances of the Geospatial Web. Like the Internet, which is composed of many local extranets and intranets, the Geospatial Web rests on a common framework of open standards and standards-based technologies. The importance of such open platforms is firmly established (Cargill 1997). This discussion provides a clear description of the Geospatial Web and its role. Key standards and capabilities are highlighted. The notion of these standards-based technologies as an *infrastructure for the Geospatial Web* is introduced. Specific examples of real-world applications being deployed on the Geospatial Web using this infrastructure are discussed.

At the outset there needs to be a clear differentiation between geographic data and map images. While maps may be the most commonly recognized product associated with geospatial data and applications, they are just that: *one product at the end of a long supply chain*. This supply chain acquires and fabricates geospatial data to develop new information, to make decisions, to assist in a broad-range of modeling and simulation and for many other purposes, one of which is to create a map. Existing standards such as WMS support the reliable delivery of map images in an interoperable framework. As such they are a component in the fabric of the Geospatial

Web. However, the realization of the Geospatial Web requires a much richer set of features distributed over a broader, interjurisdictional audience. It is this richer set of features, the standards that support the requirements implied by these features and this broader audience that occupy the remainder of this discussion.

2.2 Organizations, Integration and Data

Critical to the idea of data integration across jurisdictional or administrative boundaries is the recognition that business processes within each jurisdiction or administrative unit are fundamentally autonomous. While there might be changes in corporate or government departmental organization, our assumption is fundamentally that each organization acquires, analyzes and deploys geographic and geographically related information in order to deal with a business issue confronting that organization. Notions of information sharing that depend on new cross-organizational business processes or that demand the integration of such processes are, in our view, doomed to failure, since they conflict with the basic needs of the organization itself. Cross-organizational integration and information sharing can only succeed if they are accommodated within and transparent to the core business processes of the organization itself. Only when information sharing is achieved on the basis of such organizational autonomy should we consider integration that depends on integrated business processes. This leads to two basic premises regarding the use of geographic information in and between organizations: (i) the acquisition and use of geographic information is driven by real-world business problems; (ii) new business processes will not succeed if they are created solely to support artificial cross-organizational integration (integration for the sake of integration).

2.2.1 Primary Resources and Objects of Interest

If we begin with the principle of organizational autonomy expressed above, we must consider the persistent information stores that support the organization's business process(es) as the primary resource. Given this, the applications that update, process and display that information (underwriting organizational decision making) constitute an essential secondary class of resource. The persistent information stores contain the objects of primary interest to the organization. These objects are modeled and represented in ways that address the organization's immediate and long-term concerns and objectives. This basic set of relationships and operational dynamics is independent of organizational size and exists in application domains that include resource exploitation, transportation security, environmental protection, or the planning or operation of new urban infrastructure.

These primary resources or objects of interest can be very dynamic. They might be generated or modified by an organization as a result of its core business processes; e.g., tax parcels, property ownership and the location and condition of mobile assets. Primary resources can also be effectively static, seldom updated but providing critical input for data processing and decision making. For instance, a police department will clearly want to update objects that relate to crime incidents, traffic accidents or the whereabouts of serious criminals. However, it is unlikely that they would be interested in recapturing the layout of buildings or roadways. (Note, however, that the police may close or block a road or deny access to a building). The same police department will want to have access to additional information that they do not "own" or update such as the location of fires, major civic events and other items that could become of concern to the police. In a similar manner, the fire de-

partment has a primary interest in the location of fires and hazardous materials and would be responsible for the update of these objects and would have an interest – but not an update interest – in objects such as the location of roads, buildings and large civic events. In the former case the fire department acts as the primary custodian and/or author or authority for specific data (e.g., locations of fires, profiles and locations of hazardous materials), while in the latter instance the fire department's role is primarily that of a consumer, incorporating relevant data produced by other organizations (e.g., road networks, address ranges, building footprints and construction materials) into models and workflows to improve decisions.

One should also note that objects updated by one organization may significantly impact the business processes of other organizations. For example, the occurrence of a large fire (primary interest of the fire department) will quickly become of interest to the police or other traffic management organizations.

2.2.2 Objects and Roles

While these issues have been expressed in terms of organizations, however, they may be more accurately considered in terms of roles. Roles are usually mapped to a particular organization, but this need not be the case in all circumstances. In most application domains we will recognize the notion of an authorized observer, meaning someone who can report the existence of or change in some object, which they are not primarily responsible for, the reporting typically taking place to the responsible organization. Thus, we have the common citizen able to report a fire to the fire department or an emergency of some sort to a centralized 911 service or contacting the police to report a traffic accident.

2.2.3 Standards for the Geospatial Web

The Geospatial Web is ultimately enabled by widely adopted, open standards. These standards emerge primarily from the mainstream Web communities (XML, W3C, etc.) and from industry consortia that focus on specific functional areas or topical areas. For instance, the OASIS Consortium concentrates on Registries and Registry Services in its specification (Fuger et al. 2005). The Open Geospatial Consortium (OGC) has had a committed global membership that has worked for more than 10 years to specify a framework supporting interoperability in geographic applications. To underscore the global nature of these initiatives, many of the specifications are rationalized under the umbrella of the International Standards Organization (ISO).

With this background in hand, we can now move on to the consideration of the key standards and technologies that enable information sharing and deliver infrastructure for the Geospatial Web.

2.3 GML: A Lingua Franca for the Geospatial Web

A common language capable of expressing geographic information is required to enable information to be shared in the various ways discussed above. The Geography Markup Language (GML) provides such a language (Lake et al. 2004).[28] As such, GML provides a fundamental infrastructure that enables the Geospatial Web.

2.3.1 GML Basics

GML is an XML language for the encoding of geographic and geographically related information. While it includes geographic information, actually any information can be encoded in GML, and there is no requirement that the information be related to location or time.

GML is written in XML Schema, which delivers inherent extensibility to GML. Users of GML create their own object vocabulary (i.e., object types) by writing GML Application Schemas. These XML Schemas make use of GML schema components (e.g., time, geometry, etc.) and follow simple, structural rules for GML. The ability of users to create more or less any object in GML is essential. It is this capacity that enables GML to function both as an information transport and as a means of exposing the information model of a chosen persistent store. These capabilities are required to support information requests and transactions.

2.3.2 Type Definitions and Encoding

GML can be viewed either in terms of type definitions (schema components) or in terms of the instance data itself. Which one is more important depends on the specific application (Galdos Systems 2003). For example, if a Web service is being specified, GML can be used to define the types of the arguments in the inputs and outputs of the Web service, even if GML is not used in the actual data transport.

GML can also be viewed as an XML encoding of an extended E-R diagram, representing entities and their attributes, attribute inheritance and relationships between entities. Related entities (or classes) can be related (associated) regardless of their relative location. For example, a road object can say that it crosses a bridge object even when the respective objects are located in different spatial databases and belong to different organizations. In GML this is stated in terms of the Object-Property-Value rule, as illustrated in Figure 2.1.

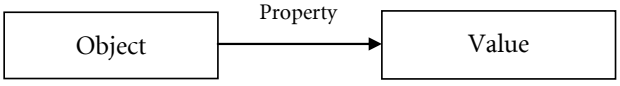


Figure 2.1: GML object-property-value relationship

The GML Object-Property-Value rule determines the structure of a GML (XML) instance document. GML Objects are those whose content model derives from a content model in the GML core schemas (GML namespace). Their children are always properties of the GML Object, and such properties can be either attributes or express relationships between objects (e.g., if the value is a GML object, the property expresses an object relationship).

2.4 Working with GML

GML was devised in order to support a fine-grained, feature-based relationship between geographic (and other) databases. GML is specified to support both transactions and requests. GML is not simply another file encoding that supports ad hoc file exchange (Galdos Systems 2003; Lake et al. 2004).

GML provides a rich collection of primitives for the encoding of time, geometry, topology, coordinate reference systems, units of measure, map styling, observations, coverages and general geographic features. This means that a GML encoding can

readily incorporate vector geographic features, imagery and real-time sensor data. This is a considerable advantage relative to having separate grammars for each or also having different Web services in each case.

Note that GML is not a format in the same sense as Shape, defined initially by the Environmental Systems Research Institute. GML does not specify the structure of any files in which the data may be transported. Character encoding and file encoding are handled at the level of the parser or a similar tool. This means that GML-based applications are independent of file structures, and appropriately written application code can be readily made independent of the details of the XML structure. It is in this independence that the real power and the real potential of GML exist.

2.4.1 Web Feature Service: A Platform for Database Integration

The Web Feature Service (WFS) protocol works with GML to provide a database and vendor neutral mechanism for both data query (data requests) and data transactions. The WFS, like GML, supports requests and transactions that can include both geographic and nongeographic information. In typical applications there are a number of ways in which WFS might be used, including (i) as database "glue" that *enables* the *synchronization* of databases, (ii) as a data service that supports *map visualization*, (iii) as a data service that directly supports *end-user applications*.

2.4.1.1 WFS Basics

The WFS specification provides a means for expressing data requests (queries) and data transactions using GML and the Filter Expression grammar associated with WFS. Data transactions are expressed at the GML feature level and include support for feature insert, feature deletion, update and locking, independent of the underlying data store. This facilitates basic communication with underlying data stores and commercial databases that have increasingly gravitated towards support for spatial data (Egenhofer and Herring 1995; Stonebraker 1996). To date, commercial implementations of WFS exist over a variety of data stores including conventional relational databases, object-relational databases and XML databases.

WFS requests using the Filter Expression (see Figure 2.2) enable the request of both spatial and nonspatial data (returned as GML) instances that satisfy user-specified spatial and nonspatial constraints. For example, one might request the Buoy features, whose location lies within a user-specified polygon and which have a "satellite transponder" data access type.

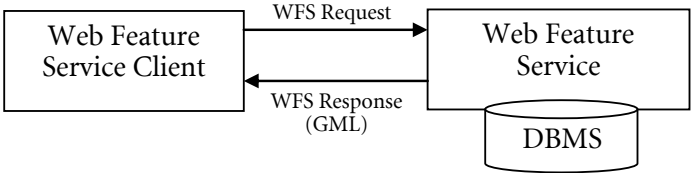


Figure 2.2: WFS and WFS client (requests)

The WFS implementation determines how the interface to the database is handled. The WFS client should not be able to determine the nature of the underlying DBMS used by the WFS. Note that the WFS client in Figure 2.2 could easily be connected to a human being, a mapping service, another WFS or an analytical (data

transformation) service or other application. It simply does not matter what the client is as long as it adheres to the WFS protocol. Example WFS transactions are illustrated in Figure 2.3.

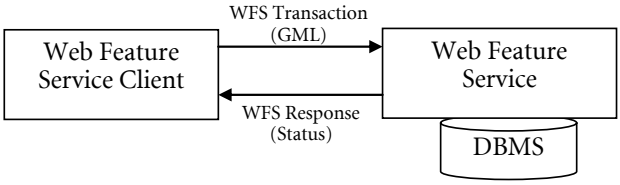


Figure 2.3: WFS and WFS client (transactions)

The WFS implementation determines how interaction with the database is accomplished. Specifics of the underlying database are transparent to the client. GML is used to encode the data and to express the transaction requested (e.g., feature INSERT, feature UPDATE, etc.; Figure 2.3). The response to the client is also expressed in GML (Figure 2.2). In this way, GML provides the basic infrastructure for interoperability across multiple, heterogeneous databases (see Figure 2.4).

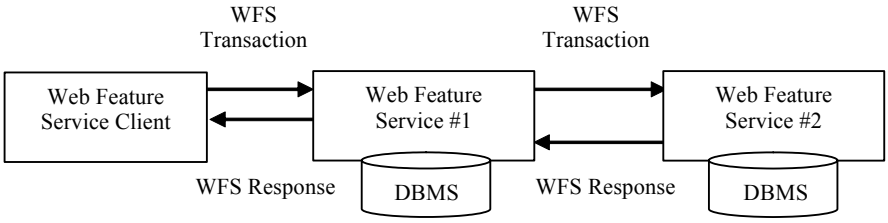


Figure 2.4: WFS transactions for database synchronization

In Figure 2.4, a WFS transaction arriving at WFS #1 is propagated to WFS #2. Assuming that these databases were in synch (for the respective feature types) prior to the transaction, they will remain so after the transaction has completed at each WFS, thus keeping them in synch.

The mechanism illustrated in Figure 2.4 *lays the foundation for publication-subscription infrastructure between geospatial databases regardless of the underlying database platform*. It means that we can readily deploy systems in which the fire department's database subscribes to changes for specific types of features from City Hall (e.g., street centerlines, building plans, etc.) and can likewise support subscribers to its data such as the police department and the water company.

2.4.2 Feature Portrayal Service

The now well-known Web Map Service (e.g., University of Minnesota Map Server)[29] provides standard interfaces to request a map. Users can control the data "layers" that appear in the map, their z-order and certain aspects of the layer appearance. While map layers (visual components of the map) are advertised by the map service, the underlying feature model is not exposed. This restricts the nature of the maps that can be produced, introduces visualization problems in integrating across jurisdictions and may cause ambiguity when the user is able to access the underlying data on which the visualizations are based. It is also difficult if not impossible to

provide for fine-grained map styling using the WMS interface, even when using Styled Layer Descriptors.

These problems can be circumvented to a very large degree through the use of a variant of the WMS, called the Feature Portrayal Service (FPS). An FPS always obtains its data from one or more Web Feature Services. Additionally, since it is a WMS, it may also cascade to other, more conventional WMS, possibly to use their layers as a background. The basic architecture of the FPS is shown in Figure 2.5.

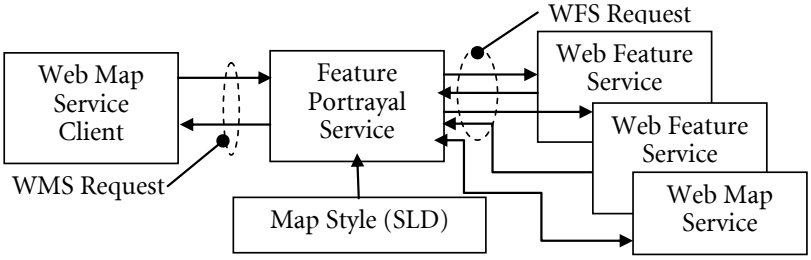


Figure 2.5: Feature portrayal service architecture showing support for WMS requests

WMS requests typically include or reference a Styled Layer Descriptor (SLD) document that contains the styling rules for treating the GML features obtained from the one or more WFS either statically known to the FPS or selected in the SLD documents. Additionally, map layers may be requested from one or more WMS, either of the conventional or FPS variety.

Note that to construct the styling rules (SLD) requires access to the WFS schemas, which can be obtained from the WFS using the standard WFS interfaces. A map style (SLD) can thus be constructed interactively by portraying the schema together with a suitable symbol library and having the user select the desired association of one with the other.

The SLD is an XML document that provides styling rules. These rules are applied to the GML features returned from a WFS to generate the visual presentation of these features in a way that enhances or increases the information delivered via the rendered map. A rule might, for example, determine that a road with two lanes be drawn as a 3-pt black line while a road with four or more lanes be drawn as a 5-pt red line. One can readily see the need to manage such rules, and to do so in relation to GML Application Schemas (e.g., style rules applicable to roads do not work for rivers). In fact, a little reflection on the interfaces and standards already introduced will readily reveal the need to manage a number of artifact types and their relationships to one another in this way. For instance:

- GML Application Schemas and the WFS interfaces that support them;
- GML Application Schemas and suitable styling documents (SLD);
- WFS interfaces and the description of the data they support;
- Coordinate Reference System (CRS) definitions and the CRS supported by WFS;
- SLD documents and related symbol definitions;
- WFS and the organizations that deploy them;
- SLD documents and the application types to which they are suited.

One might easily construct a much longer list. However, it is clear that a successful Geospatial Web clearly must be able to manage a wide range of such artifacts and do so in relation to one another (e.g., manage associations). In other words, there must be a flexible, performant and reliable infrastructure to maintain and manage relationships and associations. To accomplish this, the Geospatial Web demands a richer notion of cataloguing than was the case in the conventional world of GIS. This need is met by the ebRIM-based registry service (WRS).

2.4.3 Web Registry Service: Managing Geospatial Web Artifacts

The Web Registry Service is an outgrowth of both the OGC Catalog (OGC 2006a) and the OASIS ebRIM or e-business Registry Information Model (Fuger et al. 2006; Farrukh 2006) and builds on the strengths of each. The OGC Catalog provides a simple grammar for requests and transactions similar to that of the Web Feature Service (WFS), while the ebRIM provides a general meta-model for the description of geographic artifacts (see Figure 2.6). This ebRIM meta-model has been designated as the preferred (cataloguing) meta-model for future OGC CSW Catalog Application Profiles.

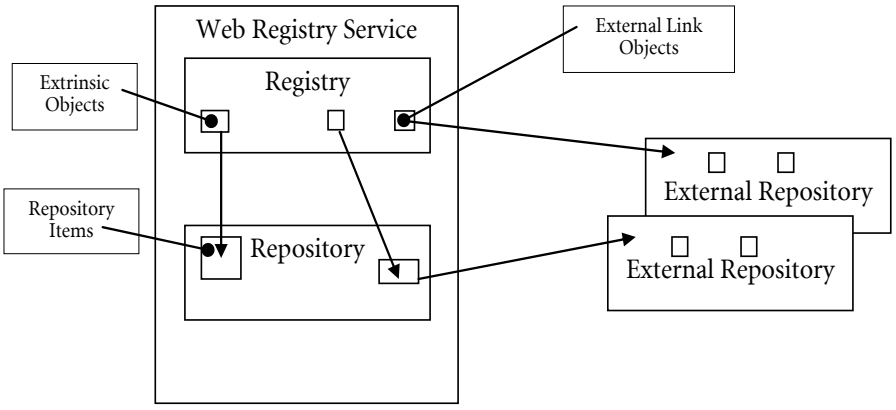


Figure 2.6: Web Registry Service architecture

2.4.3.1 Key Attributes of the Web Registry Service

The Web Registry Service includes a number of key attributes, including the following:

- the notion of *Registry and Repository*. The Registry plays the role of the card catalog in relation to the "books" (resources) in the Repository. The Repository can be internal or external, and there can be any number of external repositories.
- specific kinds of Registry objects, called *ExtrinsicObjects and ExternalLinks,* act as "proxies" for the Repository Items they reference. This relationship is a managed one in the case of the Internal Repository (it is under the control of the WRS) and a nonmanaged one for arbitary external resources that can live in any of enumerable external repositories, which are not under the control of the WRS.

- the Internal or Local Repository may indirectly reference other resources in external repositories. One might have, for example, an internal repository item that describes an image resource with the actual image residing in an external image server.
- Classification Schemes (taxonomies) and associations are used to relate ExtrinsicObjects and ExternalLinks and by reference provide classifications and associations for the resources to which they refer (e.g., the key functionality of the ebRIM model applies to the Registry).

2.4.3.2 ebRIM Infrastructure for the Geospatial Web

In addition to the notion of Extrinsic and ExternalLink objects, the ebRIM provides a range of infrastructure to support the Geospatial Web, including:

- Classification Schemes or user-defined taxonomies that can be employed to classify ExtrinsicObjects and by reference the repository items to which they refer. Taxonomies can be *used to help users discover resources of interest* (e.g., search for a Web service by its classification) and can add an additional layer of meaning to the resources they classify (e.g., classify GML feature objects using a feature-type hierarchy).
- associations are named (and weakly typed) objects that connect Extrinsic- Objects to one another. For example, a Geospatial Web might define the following associations: (i) *serves* association linking a WFS and a data set description; (ii) *servedBy* association linking a data-set description and a WFS; (iii) *holds* association linking a WFS to a GML Application Schema.
- packages are named objects that collect related ExtrinsicObjects, associations, slots, Classification Schemes and Stored Queries and are useful for a particular application domain. The contents of a package are discoverable as is the complete set of packages that are supported by a given WRS instance.
- slots provide a simple extensibility mechanism and allow a sort of typing for RegistryObjects including ExtrinsicObjects, Classification Schemes, associations and ExternalLinks. A slot is a mechanism to add arbitrary property values to any RegistryObject instance.

The Web Registry Service provides the specific OGC transaction and request interfaces for loading, removing and supporting one or more ebRIM packages on a WRS instance. Support entails the ability to query metadata for a specific ebRIM package including the ability to query the content of the associated local repository items using XPath expressions. A Web Registry Service requires the deployment of an ebRIM Package. A number of such packages will be defined and standardized by the OGC and other bodies in the near term. A Basic Extension Package is part of the base WRS specification. This package includes service Profiles for Extrinsic Object Type Definitions, Classification Schemes, Association Objects and Stored Queries.

Additional Extension Packages are currently being defined for ebRIM and will likely be subsequently standardized by the OGC and associated organizations. These additional Extension Packages include Earth Observation Products (Remote Sensing), Coordinate Reference Systems, Feature Type Dictionaries, Feature Catalogs and Units of Measurement.

2.4.4 From Standards to Technology

The standards outlined above provide the technical foundation for the Geospatial Web. To implement and sustain the Geospatial Web, three things are required: (i) technology vendors must build the component technology specified by the standards; (ii) system integrators must use the components fabricated by the technology vendors; (iii) user organizations must understand the benefits of the Geospatial Web and must be committed to an environment rich with information sharing.

Clearly, users are the direct beneficiaries of the Geospatial Web.

It is estimated that more than 50 percent of all geodata collected is collected more than once (U.S. General Accounting Office 2003). In other words, 5 of every 10 elements of geometry, image-based coverages, LIDAR generated point-cloud data and the metadata and attribution ascribed to these data elements are collected more than once, often by agencies and departments that might benefit significantly from cooperation and collaboration. The cost associated with this redundant collection of geodata is staggering when you consider that estimates indicate as much as 80 percent of the total cost of building a GIS system is directly tied to the collection and maintenance of data (CTG 2001). Together, these reports suggest that *up to 40 percent of all funds directed towards GIS could be recovered if better infrastructure for sharing and tracking data were in place.* Telephone companies map the streets. Gas companies note the location of water mains. Environmental management organizations map the same forest cover mapped by commercial forestry companies, and on and on. With a functioning Geospatial Web, much of this redundant data collection could be eliminated, reducing costs, increasing efficiencies and perhaps most importantly improving the quality of information leading to decisions.

Data changes "on the ground" are for the most part caused or detected as the result of existing business processes. Developers apply for permits to build roads and must submit their location and design to the permitting organizations before and after they are completed. A land parcel is first subdivided and registered with the land registry. The temperature distribution over a province or a county is determined by a weather reporting agency or by a collection of local sensors. The water company determines the location of the new water main before it is utilized by City Hall or the fire department. Public Health officials want to monitor waterborne bacteria and flora in relation to water treatment plant outlets. The Geospatial Web will link the various business processes together at the data level via highly synchronized databases (see WFS above) to insure that each party has access to the most current information – as it happens.

Since the "pain" of not having integrated access to information is often a shared pain, action to correct the problem is often less direct than one might hope. Two broad categories of Geospatial Web deployment can help to illustrate this issue.

The simplest case is that of a major data aggregator. In this case, a private company or government agency pulls data together, for redistribution or to provide data-driven services to end users (e.g., routing, directions, school zone, real estate, tax parcel or voting district information). Today data are assembled using a combination of manual and automated processes. Some data may be updated more or less automatically (e.g., house prices). However, other changes are only handled through bulk data purchases and manual data assembly. Absorbing and publishing these changes has a much longer turnaround time (perhaps a year or longer). Data that require manual intervention to assemble also have higher frequency of error resulting from simple, manual mistakes. The Geospatial Web will reduce the publication

time period to days or perhaps hours and reduce or eliminate the need for manual intervention. Aggregators can subscribe to features (feature changes and new features) that are published by data suppliers. The subscription-publication process can be accomplished using WFS and GML technologies. Aggregators can inspect the data sources using FPS and register their schemas (feature types), appropriate presentation styles, etc. using Web Registry Services.

A more complex case, and also a more general one, may be termed "Geographic Communities". In this context there is no single aggregator, just a collection of peer companies, government agencies and NGOs whose normal business practices depend on obtaining information from one another. Using Geospatial Web technologies (GML and WFS), publication-subscription relationships can be established between the databases of the participants, specifying access policies and the features available for publication, etc. Participants can be registered. All of this can take place with little or no changes to the underlying database and GIS application technologies employed by each of the participants.

2.4.5 Real-World Use Cases

The Geospatial Web is a vital, growing phenomenon today, not a hypothetical, conceptual construct. To reinforce this notion we provide three specific examples of how the infrastructure technology described in this discussion has been deployed to meet real-world problems.

2.4.5.1 Mission-Critical Security Applications

Open-standards WFS have been deployed to absorb the output from an array containing thousands of sensors. The WFS infrastructure is then used to disseminate these data across a large, heterogeneous community of users and analysts. Thousands of transactions occur each minute, yielding a real-time, operational understanding of conditions on the ground, providing the basis for making informed decisions in regard to critical homeland security issues.

2.4.5.2 Energy and Commerce

Open-standards Registry Services (ebRIM-based) are being deployed to support infrastructure data required by the Oil and Gas Producers Consortium, or OGP. In this case, the Registry Services are being used to manage critical coordinate reference system metadata across a highly distributed network. In addition, the registry helps manage publication/subscription relationships that help ensure that public data are made available while private data remain secure and privately held, protecting the IP and competitive advantages of individual members.

2.4.5.3 A Working SDI

A combination of WFS and Registry Services is being used to create an operational Spatial Data Infrastructure (SDI) in the capital region of a large Middle Eastern country. In this case, publication and subscription relationships across multiple municipal agencies (Fire, Water, Roads, etc.) are being managed by rules articulated in the Registry. The right-to-use authority of specific agencies in regard to specific data is sustained within the SDI. For instance, the highway department has the exclusive right to update and distribute road data. Other agencies in the network act as

consumers but cannot make changes to or redistribute these road data. This helps protect and ensure the integrity of the data. WFS resources are used to accomplish the actual exchange of data. The fact that updates are provided across agencies at the feature level as opposed to a bulk "replacement" of existing data underscores the dramatic increase in efficiency obtained via a reliance on the open-standards infrastructure technologies in the context of the SDI.

2.5 Conclusions

The Geospatial Web relies on open standards and the infrastructure of the Internet to connect geospatial data resources and producers of geospatial data with users and decision makers. The process of building the Geospatial Web is well under way with the platform of open standards that are delivered via industry consortia vehicles such as the Open Geospatial Consortium and OASIS. As is the case in any market-driven economy, suppliers and producers will rely on demand from customers to direct and drive their investments and the production of goods and services. The level of geospatial literacy has been dramatically inflated as a by-product of the success and the visibility of initiatives such as Google Earth and Microsoft Virtual Earth. The impact of these products and the millions of user experiences per day are rapidly helping to crystallize a body of requirements that the industry must respond to. These requirements include currency, reliable access and the fusion of broad, complex data types in an accessible environment that is seamlessly integrated with the browser-based environment that reflects most people's primary computing platform. In essence, the requirements emerging from these millions of user experiences per day are demanding an operational Geospatial Web, and it is this market-based pressure that ensures the success of the Geospatial Web as an infrastructure for business and administration.

Chapter 3

Imaging on the Geospatial Web Using JPEG 2000

Michael P. Gerlek • Matthew Fleagle

Abstract. *As geospatial imagery becomes more available and more commonly in demand as an indispensable part of the geospatial community's workflows, new solutions must be found for overcoming the barriers that have marginalized image data in the past – in particular, the compression of massive image sets without loss of quality and inclusion of the geographic metadata that would make imagery "spatially aware". The relatively new JPEG 2000 standard is ideally suited as a delivery technology for geo-referenced imagery, but a few features are still required for use in the Geospatial Web, including a mechanism for representing geospatial metadata and bandwidth-aware standards for client/server interchange of image data. This article discusses JPEG 2000 and gives examples of some of the emerging technologies surrounding it – largely from the Open Geospatial Consortium (OGC) – which together make it the right imaging format for the Geospatial Web.*

3.1 Introduction

The Geospatial Web is about defining the global geospatial network and doing for geographic data what the World Wide Web originally did for textual data: making it shareable, searchable and ubiquitous. The Open Geospatial Consortium's (OGC) suite of Web services and the Geography Markup Language (GML) are among a number of initiatives helping us move towards this goal by providing semantic frameworks and protocols for interoperability that are open, scalable and extensible. For archiving, sharing and querying raster (image) data, JPEG 2000 (JP2) is poised to join them.

Aerial and satellite imagery have been a mainstay of the geospatial community for decades, and in recent years – even without considering the larger Geospatial Web context – rapid access to such high-resolution data has become an imperative. One of the problems organizations face is how to store, access and exploit the massive archives of high-resolution data that are so vital to the geospatial community.

Because digital image files are so large, maintaining imagery in its raw, uncompressed form requires immense physical storage resources, and accessing raw data requires high-bandwidth networks and large-memory workstations. A typical alternative is to store imagery in compressed form, using technologies such as that specified by the original JPEG standard. But, although JPEG versions of the images may enable faster access to lower-resolution image overviews, their quality is not suitable for analysis and exploitation work at high resolution. This often leads to the practice of storing multiple versions of each data set at different resolutions – one for browsing, one for analysis, and so on. The storage and maintenance problem gets worse, not better.

Wavelet-based image compression schemes – which can support multiple resolutions of a data set within the same file without increasing file size or access time – directly address these problems. Today, a number of groups around the world are taking steps towards bringing the advantages offered by wavelet-based compression into our geospatial networks – the nascent Geospatial Web.

Despite the industry's widespread adoption and ongoing use of proprietary formats and *de facto* standards, organizations increasingly prefer using technologies based on open standards such as JPEG 2000, the wavelet-based successor to the original JPEG. Because of its promise of interoperability and the fundamentally new capabilities that JPEG 2000 technology brings to us, it is quickly becoming an accepted file format for storing geospatial imagery. However, the broad mandate of the JP2 format prevented its specification from defining precisely how to handle geospatial metadata, which means there is work to do to bring JP2 fully into the sphere of geospatial interoperability.

In this article, we will begin with a short discussion of a few of the criteria for raster imaging on the Geospatial Web and then follow with a short summary of the advantages of JPEG 2000 as a technology built for storing, manipulating and transmitting massive image data. We will then discuss a few specific technologies for bringing JP2 into the Web: specifically, we will talk about making JP2 "spatially aware" by embedding GML in it, extending the OGC Web Coverage Service (WCS) to support JP2 and storing JP2 data natively and efficiently in spatial databases.

3.2 Imaging on the Geospatial Web

Over the years, our experiences with customers' needs for imaging support – and, indeed, our industry's experiences in building networked systems – have led us to the following claims:

- *Image data are important.* There is still a significant subset of the GIS industry focused on vector/shape/feature data that use imagery only for pretty background fills, if at all. Sometimes this is justified by the product being developed, but we still often find environments where imagery can or should be used, but the cost of carrying around heavyweight raster data is just too high.
- *Universal client/server standards and interoperability are required.* This is an obvious one, too, but one that nonetheless bears repeating as some customers – and some vendors – still don't get it. Without an open, level playing field of protocols, we would not have a Geospatial Web but rather a set of unconnected islands that cannot share data or functionality.
- *Bandwidth can't be ignored.* There are many potential nodes in the Web that are bandwidth-limited but that are also too important to ignore – disaster response and the battlefield are common cases, as are networks within developing nations. Both require high-quality raster imagery, but because of their conditions cannot support high-bandwidth network connections.
- *Neither can ad hoc networks.* We must not only consider low-bandwidth networks, but also "no-bandwidth" networks. Again looking to disaster response and battlefield environments, or perhaps to remote sensor sites, we often find cases where the network connection drops – sometimes momentarily, sometimes for long periods. In such situations, image data caching can be a valuable means to maintaining the continuity of imagery work.

- *Databases are preferable to flat files.* The paradigm of using servers full of "flat files" is starting to show its age. Such systems are harder to index, secure and maintain than other more structured systems like modern, spatially indexed databases.

In addition to these claims, we know from working with customers such as the United States Geological Survey (USGS) and United States' National Geospatial-Intelligence Agency (NGA) that the Geospatial Web also requires support for very large files (tens or hundreds of gigabytes), storage of the imagery without significant quality loss, access to multiple resolutions or overviews quickly, efficient random access into the file (to support arbitrary scene requests), and so on. It is our contention that JPEG 2000 offers a way to address these needs, in ways that traditional geospatial file formats such as GeoTIFF cannot.

3.3 JPEG 2000 Technology

The original JPEG file format, named for the Joint Photographic Experts Group, was formally standardized in the early 1990s. Although the format was state-of-the-art at the time, the needs of many industries, including the geospatial industry, have since grown to include capabilities beyond those of the original specification. In the late 1990s, an ISO committee began work on a successor standard. Part I of this new standard, covering the basic encoding technology and format, was published in 2000 (ISO 2000).

At a simple "data" level, this new format offers many of the best features found in other modern raster file formats, including support for lossless encoding, multiple bands (multi- and hyperspectral), 16 or more bits of precision, signed and unsigned data types, user-extensible metadata, and so on. Beyond that, however, JP2 also provides for high-quality compression, multiresolution representations, selective decoding and progressive transmission (Taubman and Marcellin 2002, Marcellin et al. 2000).

For geospatial image users, this means JPEG 2000 provides greater control over encoding choices that can favorably affect the storage requirements, transfer performance and usability of high-resolution data, as well as help reshape common imaging workflows.

3.3.1 Workflows

When working with JPEG 2000, one must be familiar with three basic parts of the tool chain: encoding, decoding and transcoding. Figure 3.1 shows some sample workflows using these three parts of the JPEG 2000 tool chain.

- *Encoding.* An encoder converts an image into the JP2 format. The image is most simply encoded in a *lossless* manner, meaning that no data are discarded and the original image can be perfectly reconstructed. In lossless form, the encoded file is typically half the size of the original raw image file – a 2:1 ratio – due to the nature of the JP2 data representation. Alternatively, the user may choose to allow some data to be discarded, yielding a *lossy* image with a higher compression ratio (and a smaller file size). For color RGB (red-green-blue) images, a compression ratio of 20:1 typically yields an image that, while numerically lossy, is *visually lossless* – that is, to the human eye the image is indistinguishable from the original.

- *Decoding.* A decoder converts images from the JP2 format back to raw pixels for viewing, exporting to other formats, and so forth. Because the JPEG 2000 wavelet-based technique encodes an image at multiple resolution levels (comparable to pyramid-based schemes but without the file-size overhead), the decoder can choose which levels of the image to extract: at, for example, full, half- or quarter-resolution. Unlike some compressed file formats, JP2 supports extraction of arbitrary scenes from an image. This *selective decompression* feature obviates accessing the entire image file to decode a small portion of it, which means that viewing is faster.
- *Transcoding.* A transcoder transforms a JP2 image, or a subset of a JP2 image, into another JP2 image. Old-style JPEG (.jpg) files are "static", in that you can't do much with them besides extract the entire image. With JP2, however, images can be further compressed, internally reorganized, cropped, scaled, rotated and more. These operations happen within the JP2 wavelet space, which means that there is no need for the decode/re-encode process that these types of operations require in other formats. The innovations of the JPEG 2000 transcoder are significant enough to lead not only to faster workflows but to entirely new ones.

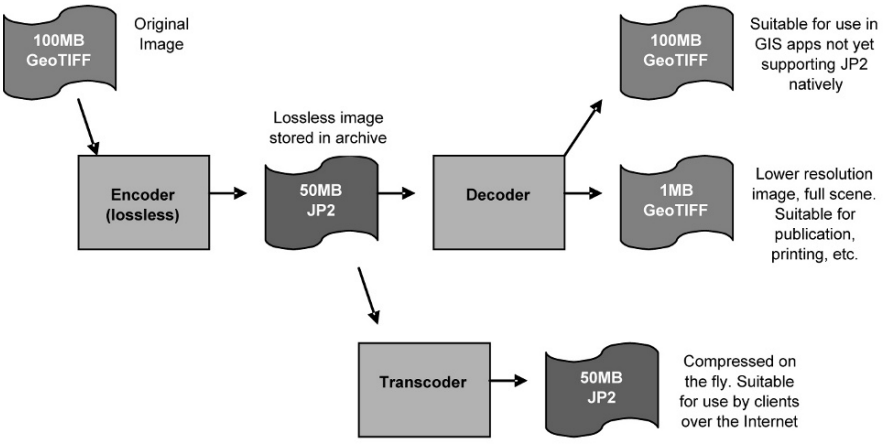


Figure 3.1: Sample JPEG 2000 workflows

3.3.2 Wavelets and Bitplanes

In simplified terms, the encoder, decoder and transcoder in the JPEG 2000 process put image data through two key transformations represented by the wavelet transform and the arithmetic encoder. Thus, during encoding, pixels of the original image are first transformed to become wavelet coefficients grouped into small "blocks". *Resolution levels* (representations of the image at different resolutions defined by width and height in pixels) are created in the wavelet transform stage.

Once in wavelet space, the data are run through an arithmetic encoder that compresses the image data losslessly and arranges them in the most efficient order for access. Lossless compression can be thought of as follows: if you or I tossed our own groceries into a brown bag until it was full and then turned the bag over to a counter

attendant, the experienced bagger could undoubtedly find a better way to pack the contents and create more room without discarding any of our items. In the same way, the arithmetic encoder identifies redundancies and other inefficiencies in the strings of bits that make up an image, resulting in a smaller file size.

It's worth noting that not all bits are equally important. The most significant bits make up the bulk of an image's visual representation, whereas the least significant bits provide the finest degree of detail. During arithmetic encoding, the data are arranged into groups called bitplanes, such that the most significant bits are all contained together in one bitplane, the next most significant bits in another, and so on down to the last bitplane containing all the least significant bits.

A JP2 file, then, is an efficiently ordered collection of bitplanes. In this lossless encoding, all the bitplanes have been retained. By comparison, lossy compression is a transcoding operation that simply discards selected bitplanes, resulting in a corresponding reduction in image quality or detail.

The decoding process is merely the wavelet transform and arithmetic encoder steps in reverse. Whatever bitplanes are present in the file are put back together into the small blocks in each of the resolution levels, which are then reconstructed into the original image. Because the JP2 file size is reduced by the removal of bitplanes and because the performance of the decoder is a function of file size, decoding of JP2 images is faster than encoding, especially at higher compression ratios. This is a huge benefit to users because images are encoded only once but decoded – that is, accessed and viewed – many times.

3.3.3 Adding Complexity

Although simple in its overall encoding process, the JPEG 2000 standard is complex when it comes to practical use, and the complexity is really an embarrassment of riches. JPEG 2000 offers many different ways to process an image – some much better than others for a given workflow.

To take one example, the order in which bitplanes are stored in a JP2 file – called the *progression order* – can significantly affect decode speed. Data at the front of the file are more readily accessed than data "further back", leading to unpredictable server response times.

Consider a progression order in which the lowest-resolution and lowest-quality data are collected at the front of the file. This would be good for workflows oriented towards panning quickly around an image at low resolution; in contrast, using this image in a workflow that involves zooming and analyzing details requires significantly more disk access and performance decreases. This issue is, at its heart, more about managing file (disk) I/O than JP2 itself – but even this can be addressed, as we will later see.

3.4 JP2 and Metadata

The JPEG 2000 committee designed the compression algorithms and file format to support a broad base of application requirements, but they did not extend the specification to support any particular application domain such as GIS or medical imaging. Perhaps wisely, the committee chose to define only the implementation for adding metadata, leaving the various communities to work out the contents. Thus, the appearance of JPEG 2000 on the geospatial imaging frontier is a bit like Hercules showing up at a rugby match and saying, "If you teach me this game I'd love to

play." The raw power and athletic ability that every team wants are present; the lad just needs to be told what the rules are.

Our first step, then, to bring JP2 into the Geospatial Web is to make the format "spatially aware"; an image needs a mechanism for describing its coordinate reference system (CRS) and geographic location. The simplest approach, which has been used for many image file formats, is to add a world file, which is a text file containing just the (x,y) position, rotation and pixel resolution. While simple, this approach has the least expressive power. More expressive is the *de facto* GeoTIFF[30] standard, which extends the well-known TIFF format by adding tags describing not just location and resolution but also the projection system being used.

Looking to provide an even better solution, groups with expertise in geo-imaging and geographic metadata have been working within the OGC to develop a modern mechanism for spatially enabling JPEG 2000.

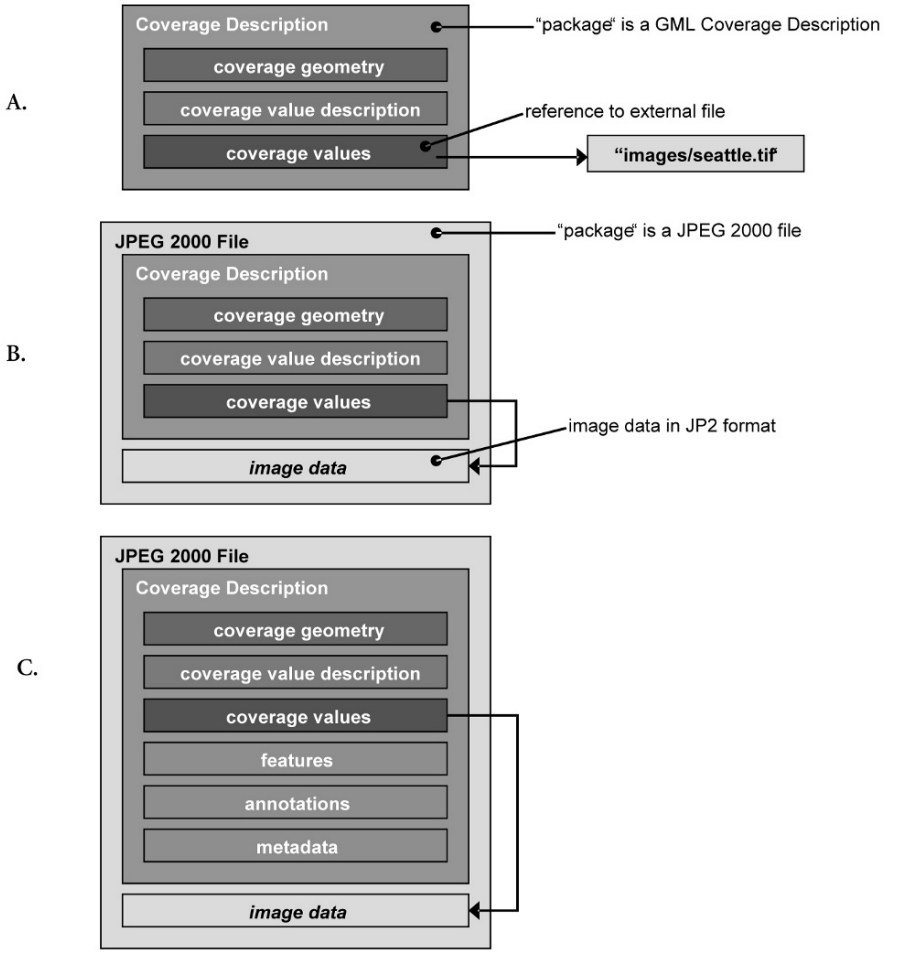


Figure 3.2: Three GML coverage description models

3.4.1 Enter GML

Defined by OGC and now an ISO standard in its own right, the Geography Markup Language (GML) is an XML-based language for representing many types of geographic content. Using GML primitives, one can describe CRSs, units of measure, features, geometries and topologies, coverages, annotations, and so on. GML does not, however, include object types such as "Road", "Political Boundary" or "Remote Sensor". In the typical usage scenario, GML is used to construct from the primitives an application schema that defines community-specific object types like these. Application schemas can be used to define what data, and what types of objects, should be used to make up a GML instance document.

In the case of raster imagery, GML defines a specific kind of feature called a *coverage* that is, among other things, a rectified, geo-referenced image. Figure 3.2a shows a coverage description using the GML model. For purposes of this article, "coverage" can be defined as an object that contains the geolocation of the image data, the type of image data and the image data itself. Note that, in this case, the image data are located in some file external to the GML description of the image.

3.4.2 A Spatially Aware Image Format: GMLJP2

The JPEG 2000 standard allows XML metadata to be stored in a number of user-specific "boxes" within the file, alongside the compressed image data. The GMLJP2 standard, published by OGC, defines a GML application schema for storing coverage information within a JP2 file (OGC 2006b). This standard specifies, in effect, where the various pieces of XML data are to be placed inside the JP2 file and what GML data they are to contain. In the minimal instance, this is just the coverage description discussed above. Figure 3.2b shows how the coverage is represented within the JP2 file, using the GMLJP2 model. Note that that coverage data now reference not an external file but the encoded JP2 data within the file itself.

With GML, of course, much more can be expressed than just the basics of the coverage offering. For example, we might wish to describe a feature, such as a runway, and annotate the feature with some information for the user of the imagery (see Figure 3.2c). With these abilities, GMLJP2 is obviously much more powerful than the GeoTIFF tags used with TIFF.

Let us consider a few examples common to many typical workflows.

3.4.2.1 Example 1—Including Coordinate Reference Systems

Figure 3.3a shows the simplest case: a JP2 file that contains the normal "image data", plus a box of "XML data" containing the GML-encoded coordinate information. Figure 3.3b shows how this is similar to the encoding information available with the GeoTIFF extensions to the TIFF standard.

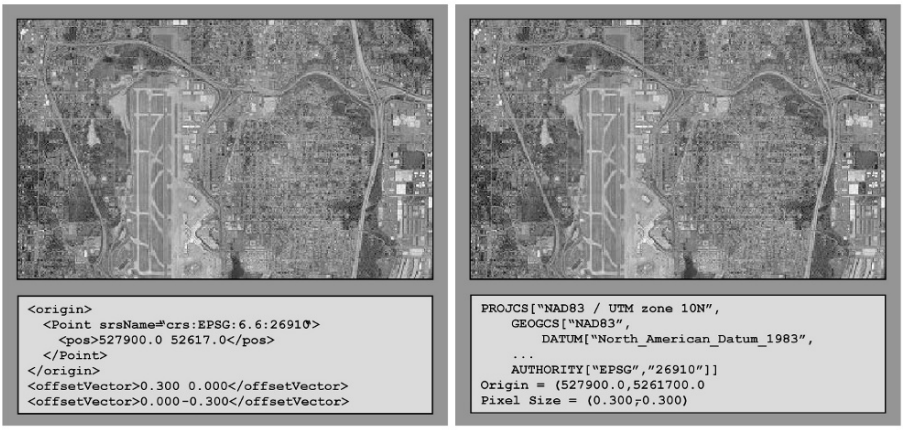


Figure 3.3: Two ways of spatially referencing a JPEG 2000 image

3.4.2.2 Example 2 – Integrating Other Useful Data

Now let's go beyond what we can do with GeoTIFF. Figure 3.4a shows an aerial image of Seattle, Washington, and Figure 3.4b shows the road network for the region. In most workflows, these data would have to be represented in two different files, in two different formats: perhaps a GeoTIFF and a Shapefile. Using GMLJP2, however, we can store the image data (with all the advantages JPEG 2000 has over GeoTIFF), add the geographic extent information as we did above and then store the road features (with all the advantages GML has over vector files); see Figure 3.4c.

Figure 3.4: Annotating images with GMLJP2

Additional information might be added to the image by some later user. A specific building within the coverage might be called out as an area of interest, for example; GML can be used to locate the building and detail some observational data about it.

The advantage of having feature data and metadata within a single file is more than just a notational convenience: we now have a complete representation of the area and all its feature data that can be treated as a kind of self-contained package, almost a "database", suitable for distribution to third parties.

3.4.2.3 Example 3 – Providing a Junction for Distribution

GML is the common language used within the OGC Web Services ecosystem, such as for Web Feature Service (WFS) data. GMLJP2 provides a ready means for image and feature data to be combined as a single output type for some downstream user. In Figure 3.5, we show how the above Seattle example might be played out using these Web services for a utility company field engineer: a WFS and a Web Coverage Service (WCS) serve up the feature and raster data, respectively, for a given region, and these data are then presented as a GMLJP2 file for later offline consumption by a custom application on a handheld device.

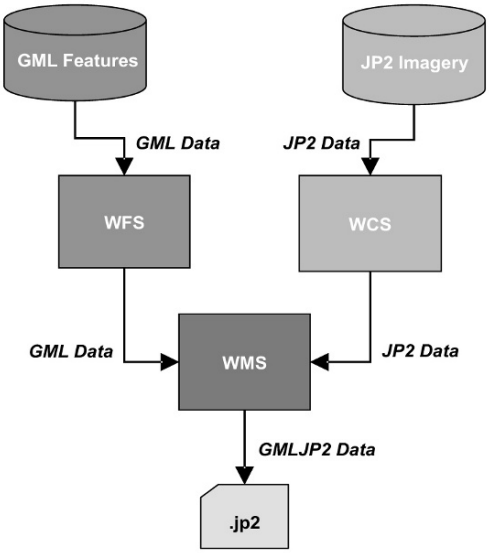


Figure 3.5: Serving GMLJP2 image data from OGC Web Services

We might also wish to provide metadata about the image, indicating perhaps the date the image was captured. We might even define an application schema to describe the sensor model being used to capture the image, or the precise positioning of the camera, or information about air quality and cloud coverage. GML annotations might also be included to indicate how a client might render this image using styling rules.

3.5 JP2 and Web Services

Because the Geospatial Web is a network of systems, we need to support imaging workflows in a client/server environment. This is only useful and interesting if done in a standards-based framework of interoperability.

OGC is building just such a set of Web services for the interchange of geospatial data (Doyle and Reed 2001), of which two are becoming well known and deployed today. The Web Map Service (WMS) is used to generate "maps" of a given area containing shape (feature) data or raster data or both, typically as JPEG or PNG images intended for human viewing. The Web Feature Service (WFS) is used to serve up GML feature data for processing and consumption by client applications.

A third service is the Web Coverage Service (WCS). The principal function of the WCS is the "get coverage" request, in which a client application requests coverage data for a specific geographic region. The server returns the coverage data in any of a variety of formats from GML to GeoTIFF.

One can easily imagine a system that combines these services: for example, a WFS server and a WCS server providing the back-end feature and raster data support for an outward-facing WMS server.

3.5.1 WCS and GMLJP2

Unfortunately, WCS does not support JPEG 2000 as one of its data return formats. Indeed, while WCS does have some of the underpinnings we'd like to take advantage of with JPEG 2000, such as the ability to request an image at a given resolution (scale), there are some capabilities that are not yet expressible with WCS, such as the ability to request a coverage data not just by spatial region but also by bits per pixel (quality). Nonetheless, such a WCS+JP2 extension to WCS is relatively straightforward to define, and indeed OGC is currently untertaking work to do just that.

This extension enables a client to take advantage of the features inherent in JPEG 2000, such as compression and multiple resolutions. With the GMLJP2 standard for embedding information in the JP2 file, the image data returned to the client can also be fully geographically described and self-contained. As an example of this, consider the previous example of combining the WFS, WCS and WMS. What if the JP2 and GMLJP2 formats were to be used across this network? With a corresponding extension to allow WMS to return not just JPEG but also JPEG 2000, the map server could provide a client with a rich file containing both the raster data and the features in the requested region.

3.5.2 WCS and JPIP

Our next criterion for imaging workflows on the Geospatial Web is about supporting bandwidth-constrained environments. Take the case of trying to pan quickly across a large data set, hunting for the particular area of the image you are interested in. Two possible "form factors" for this problem can be considered: a low-bandwidth, small-screen mobile or handheld device or, alternatively, a "normal-bandwidth", full-screen workstation using high-resolution data. In both cases, as the user pans the client is making new request after new request. Unless the application has a means of doing asynchronous updates to the screen, the user experience can be painful. Even tiling schemes, such as those used by many current Web mapping clients, improve the user experience only slightly: initially displaying a blank, gray square while waiting for the full data to be downloaded isn't as visually useful as showing an interim low-resolution version of the area.

Fortunately, one of the parts of the JPEG 2000 set of standards is a protocol for progressively streaming imagery from a client to a server (ISO 2004). The JPEG 2000 Interactive Protocol, or JPIP, uses the same data structures as the file format to incrementally deliver pieces of the image to a requesting client, with no algorithmic overhead to preprocess or transcode the image. Each of these delivered pieces is fully "indexed" in that the JPEG 2000 decoder knows "where" in the image the data belong in terms of their spatial position, resolution and quality contribution. At any point in the process of downloading the data from the server, the smart client can translate the bytes currently available to some approximation of the image. If only a small number of bytes have been transferred, the image will be of low resolution

and/or low quality; as the user "idles" over a scene, more data are transferred and the representation of the image increases in resolution and quality. Within OGC, work is going on to extend WCS to support dynamic JPIP streams as well as static JP2 image data; a possible architecture for such a system is shown in Figure 3.6.

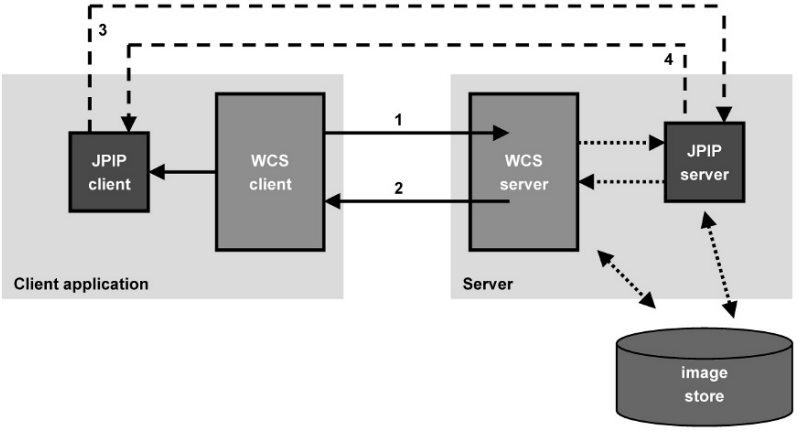


Figure 3.6: Architecture of the WCS/JPIP service

Along with the encoded image data, the JPIP protocol can be used to transmit any other information stored in the source JPEG 2000 file. Specifically, if the source follows the GMLJP2 standard, then the GML data – features, styling information, observation data – can be sent to the client as well (the JPEG 2000 and JPIP standards do not address the problem of compressing or selectively accessing XML data, however).

As an additional benefit, the image data transmitted via JPIP can be cached at any intermediate node within the network. Consider a small LAN out in the field, with one satellite dish used to periodically connect the LAN to a remote, central server. By caching the JPIP data within the LAN, each of the clients can share a single copy of as much image data as has already been downloaded, without requiring a constant network connection. The motivated reader is encouraged to consider what this potentially means on a larger scale: peer-to-peer sharing of image data, *à la* BitTorrent, but with built-in support for compression, progressive display, pyramids, etc.

3.6 JP2 and Databases

In describing a model where the data for a single image are transmitted asynchronously in discrete pieces and spread and cached across multiple systems, we are clearly getting away from the file-based model of storing pixels. Because a JPEG 2000 image can be broken down into very small, indexable constituent pieces (wavelets and bitplanes – the heart of the JPIP strategy), there is no need to keep the image data in a file at all. An alternative is to store these pieces in a spatial database, and let the database system handle the indexing and look-up work normally done at the file level.

Storing a JPEG 2000 image in a database offers three advantages. First and most obviously, databases can offer significant IT management benefits over flat file systems, such as authentication and auditing, version tracking, backups, and so on. Second, using a spatial database enables a user to store both feature data and raster

data within the same storage system, which simplifies data management. The third advantage is performance. For very large (gigabyte-sized) JP2 files, the time required to seek through the file for requested data can be substantial. Moving this burden from a file-based I/O system to a smart, caching, indexed database changes the cost equation completely. It is also worth noting that storing the image in this decomposed form within a database allows the database to act almost as a JPIP server.

As with the JPIP discussion above, where we refer to "JP2 files" we could equally well refer to "GMLJP2 files". As feature data continue to migrate into databases, we would expect such a GML-based format for interchange to become increasingly useful. Specifically, as image data become more like "regular" feature data, we can start treating them in more of a first-class fashion: querying, tagging, cross-referencing, and so on.

3.7 Looking Forward

The Geospatial Web will require support for imagery and at the same time pose new challenges for the storage and delivery of that imagery: better spatial data support, advanced client/server workflows, richer user experiences, low-quality network support, and more. Algorithmically much more complex than its predecessor, JPEG 2000 provides a solid foundation to meet these challenges, and work is well under way to prove this in real-world workflows. The components described in this article are either shipping in products today or are under active development within OGC.

Looking forward, there is still more to do in the next few years to enrich the visual content and user experience of the Web. We alluded above to the prospect of using a peer-to-peer mechanism for distributing large data sets. We also see work needed to address the real-time aspect of live feeds, the mathematical complexity of multi- and hyperspectral image processing and the continuing technical and social issues surrounding security and rights management. While JPEG 2000 may not be the right foundation for all these concerns, they will nonetheless need to be addressed to further the shared goal of a truly global, content-rich Geospatial Web.

Chapter 4
What's So Special about Spatial?

Glen Hart • Catherine Dolbear

Abstract. *Geospatial information can act as a thread that can be used to integrate information from heterogeneous sources. It does so by exploiting common location information components that often exist across different domains. As such it has the potential to be a valuable resource in the implementation of the Semantic Web. This chapter examines the challenges of adding a geospatial component to the Web, with particular reference to doing so in a way that also supports the current initiatives to semantically enable the Web. It identifies those that are largely peculiar to geography and those that, whilst issues within geography, are also likely to occur in many other domains.*

4.1 Introduction

There is a long-lived though unattributed belief that 80 percent of all information has a geographical component: by this it is meant that a significant proportion of databases contain information either directly or indirectly referenced to physical locations. Such information will include obvious candidates such as digital mapping, environmental information and planning information. It also covers information from other domains such as marketing, insurance and so on. Any information that makes reference to a postal address can be considered geospatial.

In most cases the geospatial component is not dominant; it plays a supporting role rather than being central to the business objective. An insurance company will be interested in the location of insured properties to determine flood and subsidence risk, but its fundamental interest will be in the overall risk factors and property value. The geospatial nature of the information helps the general information integration problem, as the geospatial component is often one of the most commonly shared information types between different data sets. So the insurance company is able to calculate flood risk through the intersection of property location, digital map information containing river location and perhaps meteorological information for the area.

If it is accepted that geospatial information can form an important element in information integration, it is quite apparent that it will also serve an important role in the development of a more sophisticated World Wide Web. That location is important is certainly recognized by the search engine giants: Microsoft with its Live Local, Google, Google Maps and Google Earth and Yahoo! Local. These initiatives have all been created in response to user needs for location-based information services. None is able to fully exploit the geospatial aspects of the information being searched, because of the Web's weak representation of location. Another weakness is the lack of semantics: not only is the Web unable to fully appreciate the spatial relationships that exist between the cities of Southampton and Portsmouth, for example, but it also lacks understanding of what a city is in the first place. Thus, there are strong arguments for both semantically and geospatially enabling the Web. And indeed, the role of the Semantic Web as a vehicle for information integration is ex-

plicitly recognized by its creator: Tim Berners-Lee (Shadbolt et al. 2006). This extends beyond just Web pages. Commenting on the future of the Semantic Web, Tim Berners-Lee recently asserted that we "need to look at existing databases and the data in them" (Runciman 2006). Although research interest in this "deep Web" is increasing, the process of linking an ontology to a legacy relational database raises many semantic issues that have, to date, largely been ignored. Of particular interest to Ordnance Survey, the national mapping agency of Great Britain, are the case of mapping an ontology to a spatial database and how to combine spatial and description logic queries and modeling paradigms for efficient performance.

This chapter examines the challenges of adding a geospatial component to the Semantic Web. It identifies issues that are largely peculiar to geography as well as those that occur not only in geography but also in many other domains.

4.2 The Geospatially Peculiar

The geospatial domain shares many issues with other domains when considering how to represent and exploit domain knowledge using semantic Web technologies. However, certain issues are far more relevant to the geospatial domain than others (although many may be shared with the more general spatial domain). These peculiarities are examined in this section.

4.2.1 Spatial Relationships

Given that what makes geospatial information "different" is the concept of location, it is very obvious that spatial relationships between objects that have location are very important. Central to all Geographic Information Systems (GIS), and more lately to spatially enabled databases systems, is the support of various spatial operators designed to determine the relationship between geospatial objects. For example, all will support (under various names) algorithms to determine containment, overlap, (spatial) disjointness and so on. Significant work has been done to formalize these relationships, such as the 9 intersection model (Shariff et al. 1998) and Regional Connection Calculus (RCC8) (Randell et al. 1992).

Due to their formal expression, these calculi can be expressed as topological relationships using description logics (W3C 2004) and hence in languages such as the Web Ontology Language, enabling reasoners to perform inference over spatial information based on topology. As an example, RCC8 has been implemented using OWL to support an experimental ontology editor (Smart et al. 2006). However, geospatial information rarely contains explicit topological information. It is more usual for these systems to determine specific topological relationships through geometrical calculation on information with positional information. For example, it is unlikely for a geospatial database to explicitly contain the topological relationship that a specific house is "contained within" a specific garden. Rather the containment operator will be used to test whether the building is indeed contained within the garden. Such approaches are necessary because it would be impractical to explicitly precompute all topological relationships between all the objects in a geospatial database. However, this contingency means that existing reasoners are unable to perform topological inference on the majority of geospatial information, as they are unable to compute the necessary topology. Furthermore, as Lemmens et al. (2006) have identified, relationships may be indirectly expressed through other means such as postal addresses. Thus, reasoning is limited to qualitative reasoning over the ontology rather than quantitative reasoning over instances held in the database. Cur-

rently there are investigations into the possibility of separating spatial queries into spatial and aspatial components (Dolbear and Hart 2006). The spatial component is executed first and the results set is presented to the reasoner, which completes the query. Although a pragmatic solution given the current technologies, such solutions can never be complete answers since the reasoner may discover the need to investigate further spatial relationships. To arrive at a complete solution, there are two possible paths. Either languages such as OWL will need to be modified to explicitly support spatial relationships and will in turn need to be supported by appropriately enhanced reasoners or the reasoners will need to be modified to enable certain class properties to be mapped to database functions or Web services. This will not only enable the mapping of spatial properties but also other properties too, enabling a more general solution to be arrived at.

4.2.2 Vague and Uncertain Location

The spatial relations described by the 9 intersection model, RCC8 and those implemented within spatial databases and GIS are based around "crisp" geometry. It can be precisely determined that a house is or is not contained within a garden because both have a precise (or crisp) geometry and location. However, this is not true of all geospatial objects. Consider an area such as the Lake District in the United Kingdom; it has no well-defined boundary, and as a result it is a vague object. It is possible to know with certainty that some things are within the Lake District such as Windermere because as a lake it is one of the things that define the Lake District. Precisely where the Lake District ends is uncertain because no one has ever defined a boundary. Imprecision in recording geospatial information can also result in uncertainty. Consider road traffic accident information. Typically this will be recorded against particular stretches of road but will not record precisely where individual accidents took place. Thus, although an accident (a spatial event) took place at a precise position along the road, the recorded information cannot tell us where this was. We can only know that it took place on a certain stretch of road and are uncertain about exactly where.

Existing geospatial implementations enforce some form of crispness on information that is vague or uncertain. Artificial boundaries may be created for vague objects, and road traffic statistics may be treated as if they are point objects. These are no proper solutions; merely work-rounds given the limitations of existing technology. Attention has been directed towards developing better solutions to these problems, for example, super-valuation semantics (Cohn et al. 1997) have been applied by Bennett to represent the vagueness that is implicit in the notion of a forest (Bennett 2001). Although boundaries are created, super-valuation semantics differentiates between areas that are known to be part of a vague object such as a forest, those areas that are definitely not part of the forest and those areas that might be. Furthermore, using super-valuation semantics, it is possible to delay making a decision as to where these boundaries might be until a user's context is known.

Another approach is *anchor theory* (Galton and Hood 2005), which allows the query results to transmit the notion of uncertainty rather than enforce some arbitrary boundary or collapse information to a point. For example, say we have information on a road accident that occurred along a stretch of road that runs across the counties of Devon and Somerset but do not know precisely where it occurred. Anchor theory represents this with an "anchored in" relationship. That is, the location of accident is "anchored somewhere within" the road stretch. A spatial query asking if the accident occurred in England would produce the answer "yes" and within

Wales "no" since it is known that the road stretch lies within Somerset and Devon, both of which are within England. If asked whether the accident occurred in Somerset, the answer would be "maybe", thus preserving the uncertainty.

As with the crisp spatial relationships, relationships to handle vagueness and uncertainty, such as those indicated above, are also required and need to be implemented as a fundamental part of any semantic technology.

4.2.3 Vague Relations

The discussion so far has concentrated on the nature of geospatial objects in terms of the crispness of their location. However, certain spatial relationships can themselves have components that give them vague properties. Consider the relation "near". Whether something is near or not depends on the context in which the question is asked and on the nature of the objects being compared. The scale of distance and thus what is meant by "near" will vary enormously: consider the distances that are judged appropriate when comparing the results of a question such as "which pubs are near my house?" with "which airports are near London?" In the first case, "near" typically refers to a distance of a mile or two; in the latter, airports within 30, 40 and perhaps 50 miles of London might all be considered near.

At present no solution to this problem exists. Indeed, in an absolute sense there can be no perfectly correct solution. Some very elaborate solutions have been devised (Gahegan 1995). Such proposals have tended to be overly complex in terms of the number of contextual factors taken into account, which renders practical solutions impossible to implement, as it is never possible to accurately quantify these factors. Dolbear is investigating the development of a much simpler algorithm that will attempt to approximate the results of human reasoning (Dolbear and Hart 2006). The approach is to determine nearness through a combination of object footprint size and frequency of occurrence. The conjecture is that these will serve as measures that can determine the likely range for which "near" will be deemed to operate. Footprint size is assumed to be proportionate to distance and population density inversely so.

A relationship like "next to" may be equally problematic. Although both the 9 intersection model and RCC8 have well-defined and predictable "next to" relationships, a "badly behaved brother" also exists. Given a situation where a house is surrounded by a garden, which in turn is physically next to a foot path running by the side of a road: if asked, "Is the house next to the road?" most people would answer "yes". A GIS would say no, since the house is not *physically* located next to the road. Thus, a version of "next to" exists where geospatial objects that are deemed to be insignificant are filtered out.

As with "near", it is likely that an algorithm can be devised that will approximate the results of human thought. A possible solution is to filter out objects that are typically considered less important or are physically less significant. In effect an importance hierarchy could be constructed that would place house and building higher in the structure than gardens, fences and paths.

4.3 Shared Issues

Whilst not attempting to be a comprehensive identification of geospatial issues that are (or are likely to be) shared with other domains, this short section attempts to highlight some of the more important issues.

4.3.1 Conceptual Fuzziness

The world is full of objects that don't quite fit into neat conceptual boxes and in very many cases, particularly in the geospatial world, these are natural things. Consider rivers and streams. Both are bodies of flowing water, and indeed there is little physical distinction between them other than size. The problems are where to draw the line between their descriptions and does it matter if a line is drawn? The problems cannot be easily resolved except by enforcing what could be arbitrary distinctions. Given that these problems have existed ever since people began to differentiate between rivers and streams, they have been managed, if not solved, by local definitions being applied. Thus, context of use is an important factor and an issue for the whole Semantic Web.

4.3.2 Troublesome Homonyms

The geospatial domain is filled with homonyms. Even within a narrow domain, such as a topographical interpretation of inland hydrology, words such as "channel", "pool" and "bank" can have multiple meanings. To an extent the problem can be mitigated by the use of multiple namespaces. In our own work we have used the rdf:label annotation property to provide a label for each term corresponding to what the domain expert would normally call the concept. Homonyms share identical annotations, but these are then mapped to class names that are made are made unique through the artifice of appending a modifier. So, for example, a "pool" can be either something that is a type of pond or an area of still water in a river. In both cases the annotation will be Pond, but the class name for the latter will be Pond.inRiver. Whilst this technique enables the homonyms to be represented, the fundamentals of the issue can only be resolved if the exact context of use is known. Disambiguation can then be performed when the ontology is applied.

4.3.3 Weak Concepts

Weak concepts are concepts that are sometimes used as a means to gather together other concepts, but which are themselves not so much poorly defined as poorly thought through. They are probably best explained though example. A water body is a good example as it is often used within geographical ontologies to group together items such as rivers, lakes and ponds. The problem is that although the subclasses may be well defined, the water body itself really only exists as a means to loosely group together these concepts. In practical terms, within the domain they are typically not used. Furthermore, because the concept is itself weak, it often occurs in other domains (or even the same domain) with a slightly different usage or position in the taxonomic hierarchy. So someone else may include seas as water bodies, another person may introduce the notion of other weak concepts such as flowing (rivers, streams, lakes) and non-flowing water bodies (ponds, canals), before grouping them in turn under water body. There is a clear and obvious solution – if they are not adding anything to the ontology, don't include them. Or if they must be added, they should not be included in the structure as a superclass, but through the properties that all qualifying members share. The sub-superclass relationship can then be supported through inference. This in turn leads to both good design and more reusable ontologies and concept definitions. This example has been included as a demonstration that it is not just technology that needs to develop, but the way in which we use the technology. The desire to build complex hierarchies that contain

weak concepts (in cartography the completely useless distinction between natural and manmade is another very popular weak concept) can be quite hard to resist. But resist we must, if we wish others to reuse what we have worked so hard to produce.

There is one good reason where it is acceptable to include weak classes. This is where they are used to enable disjoint classes to be efficiently defined, in turn leading to significant performance increases in the performance of the reasoners used to interpret the ontology. In effect the use of such classes is as an optimization technique. And so we are currently left with the problem of having to balance reusability against performance. Our own method, which expresses the ontology in both a structured English form and OWL, offers some way forward. It offers the potential for the structured English to become the visible representational form developed in a way to promote reused and the OWL form acting as an assembler code compiled and optimized from the high-level structured English.

4.4 Conclusions

This chapter has identified that the most important issue involved in geospatially enabling the Semantic Web is the incorporation of spatial relationships into Semantic Web technologies. Where the geography is crisp, there are well-founded and formally defined models that may be used such as RCC8. At least some categories of vagueness and uncertainty can be managed through solutions such as super-valuation semantics and anchor theory. Super-valuation semantics are to a certain extent reliant on context of use; thus, they touch on a more general and thorny issue of how to incorporate user context into Semantic Web technologies. Within the field of ontologies, less work has been conducted on vague relationships. Whilst it is possible to implement elements such as RCC8 using the existing language constructs of OWL, this will only enable qualitative reasoning. All these relationship models will require modifications to reasoners, and potentially to languages such as OWL as well, to enable both qualitative and quantitative spatial reasoning.

The chapter has also identified certain issues such as conceptual fuzziness, homonyms and weak concepts, which, whilst frequently occurring in the geographical domain, are also likely to be common to other domains too.

Chapter 5

Conceptual Search: Incorporating Geospatial Data into Semantic Queries

William Kammersell • Mike Dean

Abstract. *Traditional queries require users to invoke specific data sources, manually integrate data across multiple sources and interpret the results. These costly operations are increased for geospatial data, which may have elaborate formats and require complex geospatial operations. Conceptual search solves these problems by leveraging semantic technologies to give a new paradigm for querying data sources. New semantic geospatial tools are also added to facilitate geospatial reasoning. Conceptual search can be implemented via a service-oriented architecture for further benefits.*

5.1 Introduction

Current search mechanisms are inadequate for users needing to query for concepts in their own vocabulary. Conceptual search provides a new approach by asking for abstract ideas that are relative to a specific user. An example of a conceptual search is, "What is my preferred C5-capable airport near Washington, DC?" This query involves several ideas that are specific to the user such as how preference is determined, the requirements of a C5 aircraft and how close the airport must be to Washington, DC. Conceptual search enables users to specify these subjective concepts by combining the results of smaller semantic queries. For instance, a user may prefer airports that are in Virginia as opposed to Maryland, are closer to DC and have more C5-capable runways. The user may also disallow all airports that are flooded or in an active Notice to Airmen (NOTAM). Each of these factors is expressed as a base semantic query that may integrate both geospatial and nongeospatial data from a variety of disparate data sources. Conceptual search thus leverages the power of the Semantic Web (Herman 2001) to allow the formation of complex queries that target a variety of data sources and to combine the results into one abstract concept.

There are three main areas where conceptual search surpasses traditional queries. First, keyword and database searches force users to express their query in the data source's vocabulary and syntax. Users must know the words used in a text data source or the schema structure for a database. This can be especially frustrating for geospatial data sources that require special formats for complex geometry inputs. Conceptual search uses translations and semantic rules to convert the user's query with special handling for geospatial data sources. These translations are written once and then used automatically to translate a user's query. Second, users must integrate data from disparate sources to get an overall result. A user's query may not be answerable by only one data source, with parts of the answer spread among many sources. Geospatial data may further require expensive geometric operations such as boundary comparisons to combine results. Conceptual search uses semantic rules and ontologies created by users and adds special geospatial rule processing to facili-

tate data-source integration by providing geometric comparisons. Finally, current queries cannot express notions like preferences. These queries ask for ideas that combine different sub-queries into an overall result. Conceptual search allows users to combine the results of geospatial semantic queries into concepts. Thus, conceptual search furthers the state of both traditional and semantic queries by integrating geospatial data throughout the query process along with providing a powerful new query paradigm.

Conceptual search has been implemented and demonstrated via research performed by Northrop Grumman TASC and BBN Technologies. All source code is available on SemWebCentral (Dean and Kolas 2005; Kolas and Kammersell 2005). The rest of this chapter is organized as follows: Section 5.2 describes how nonsemantic geospatial data sources are invoked; Section 5.3 introduces GeoSWRL, a tool for creating geospatial semantic rules to integrate data sources; Section 5.4 delves into conceptual search and how it incorporates geospatial data; and Section 5.5 discusses some further benefits of conceptual search stemming from its service-oriented architecture implementation.

5.2 Data-Source Invocation

The foundation of a conceptual search is a set of semantic data-source queries. Each of these queries asks for a subset of the results that answer the conceptual search, such as how far an airport is from DC or if the airport is addressed by a NOTAM. These semantic queries are expressed using SPARQL (Prud'hommeaux and Seaborne 2006) select-queries, which define a set of triple patterns using variables. The query then returns a set of values that are bound to each variable. The conceptual search uses these queries by manipulating each query's returned set of bound variable results.

Semantic queries allow users to express queries in their own ontology. The user's queries can then be translated via semantic rules to query Semantic Web services. This ability to translate a user's semantic query to a data-source semantic query is a key benefit of the Geospatial Web, as noted by Egenhofer (Egenhofer 2002). However, many legacy geospatial data stores do not provide Semantic Web service interfaces. The Open Geospatial Consortium has addressed the problem of inconsistent legacy data interfaces by producing a set of Web service standardizations. Although these services help data-source invocation by creating a uniform interface, they are still not semantically enabled. For example, Web Feature Services (WFS) (Vretanos 2005) provides an XML-based query language and returns query results in XML including geometries expressed in the XML-based Geography Markup Language (GML) (Cox et al. 2003). An additional layer of translation must thus occur to translate the user's query in OWL to WFS XML, and vice versa. Conceptual search uses special query handling to translate semantic queries to WFS XML queries, and XSLT to translate GML and XML to OWL. A shared geospatial ontology gives an intermediary data representation to facilitate the translations, as shown in Figure 5.1. Kolas, Hebeler and Dean (Kolas et al. 2005) provide a detailed description of this process, along with its application to the Geospatial Web.

These translations allow users to ignore the specific interfaces of the data sources they wish to use and instead focus on what they want to ask. Geospatial data sources can have very complex methods for defining geometries, making translations tedious and frustrating. Semantic rules thus save the user time by performing the translations between the geospatial ontology and the data sources for them. Users only need to define a geometry once in their ontology, and that geometry can be reused

in many specific data-source queries through different translations. Translations can also harness a data source's specialized geospatial processing by including geospatial operations in the user's queries. WFS, for example, provides a set of geospatial filters that can be included in the XML queries. Previous Semantic Web systems that involved WFS (Bernard et al. 2003, Klein et al. 2004) focused on discovering services and translating their results, not translating a query to utilize advanced WFS functionality. Users can thus send certain geospatial operations to the WFS specialized geospatial computation handling instead of doing them themselves. Conceptual search thus uses semantic technologies to allow users to specify a query once in their ontology and then translates it into specific data-source queries. Furthermore, these queries can handle complex legacy geospatial interfaces and benefit from their specialized geospatial processing.

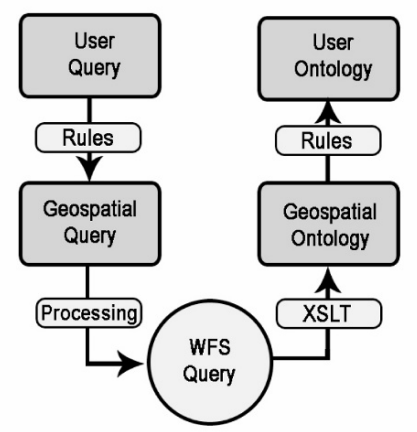


Figure 5.1: A query expressed in the user ontology is first translated into a geospatial ontology query and then into a WFS query; the results from the data source are then translated into the geospatial ontology and then into the user ontology

5.3 Data-Source Integration

Users may need to issue queries involving geospatial operations that are beyond the capabilities of a data source's processing or that combine the results of multiple data sources. Such queries have been referred to as "smart queries" by Goodwin (Goodwin 2005) for their ability to infer new information from disparate data sources. For instance, a user may query for airports that are flooded, which involves both geospatial and nongeospatial data. First, WFS services provide the locations of airports, and geospatial non-WFS services provide the locations of floodzones. Then, weather services are invoked to see if it is heavily raining at an airport. Semantic rules then determine which of the airports with rain are also in floodzones and are thus flooded. This query requires reasoning above the data-source level because it includes more than one data source and necessitates geospatial processing to determine if an airport is in a floodzone.

To facilitate the creation of geospatial semantic reasoning, conceptual search uses a set of SWRL rules (Horrock et al. 2004). SWRL defines rules as sets of patterns in the head and body of a rule. If a knowledge base contains triples matching the body of the rule, then the triples in the head of the rule can be inferred and added to the

knowledge base. A knowledge base can be a set of data-source results, using triples from multiple sources to match a body, and adding triples from the user ontology in the head. For example, data-source results can be combined by including triples from the weather service and from the airport WFS and adding triples about flooding as defined by the user ontology. SWRL also supports built-ins for additional processing, such as string manipulation and mathematical operations. A built-in can be used in the body and bind variables used in the head. SWRL thus provides a simple yet powerful means of reasoning over ontologies and knowledge bases.

GeoSWRL, a set of geospatial SWRL built-ins, was also produced with this research and is available on SemWebCentral (Kolas and Kammersell 2005). These built-ins provide basic geometric and geospatial operations that are supported by the JTS Topology Suite (JTS) (Davis 2002). JTS is developed by Vivid Solutions as a set of 2D spatial algorithms for Java. JTS requires a standardized input of polygons based on lines and points, so geospatial polygons must be converted from a data source's representation into a geospatial ontology to provide a uniform representation. The geospatial ontology entities are then converted to JTS polygons, geospatially manipulated and then translated back into the geospatial ontology. These built-ins thus provide uniform methods for applying geospatial operations to both the user and data-source ontologies.

Figure 5.2 shows two examples of using the GeoSWRL built-in DWithin. The user could create an SWRL rule defining an airport as being near a city if it is at most 100 miles from the city. This built-in is true if and only if the first argument point (airport) is within the third argument distance (100 miles) of the second argument point (city). Thus, the first rule would succeed for only the inner airport, which is within 100 miles of the city. The second example assigns the distance between the airport and the city to the distance variable. Thus, GeoSWRL provides two unique ways to incorporate geospatial operations into SWRL rules.

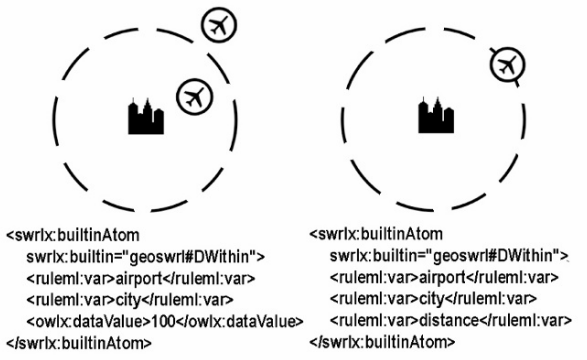


Figure 5.2: The GeoSWRL rule DWithin can be used in two ways; the first tests whether the airport is within 100 miles of the city and succeeds for the inner airport; the second assigns the distance from the city to the airport to a variable (the radius of the dashed circle)

GeoSWRL gives users the ability to integrate data-source results and to include complex geospatial operations in their queries. Users can write rules that look for triples from disparate data sources and create new triples that incorporate both. For example, a query asking for airports in floodzones would first query for airports, then for floodzones and then use the geospatial built-in DWithin to determine whether the airport is within the floodzone. If the airport is indeed within the flood-

zone, the SWRL rule will add that fact to the knowledge base. This data-source integration is easier with rules rather than new OWL classes (Klien et al. 2004) because OWL does not include geospatial definitions for classes or properties. Users can also use GeoSWRL built-ins to create geospatial operations that are beyond the capabilities of a given data source. That is, even though WFS provides basic geospatial operations, other data sources may not. GeoSWRL built-ins are also reusable and can be applied to any semantic application that uses SWRL.

5.4 Data-Source Interpretation

Conceptual search builds upon geospatial semantic queries for further expressivity and power. Conceptual search utilizes an ontology for defining the relationships between query result sets. First, conceptual search tells how result sets are combined. For instance, a conceptual search can state that the results of two queries should be unioned to produce a new result set. Thus, if a URI appears in either result set, it is in the new result set. If an airport is returned from a query for airports in Virginia or is returned from a query for airports in Maryland, then it is in a new result set representing airports on the Mid-Atlantic Coast. Similarly, an intersection of two queries requires a URI to appear in both result sets. If an airport is returned from a query for all airports in Virginia and is returned from a query for airports near DC, then it is in a new result set corresponding to Virginia's DC-Metro airports. These operations can receive many query result inputs, but can produce only one result set output. Each query result is a set of URIs, and the output of the operation is another set of URIs. These combinations can be layered to produce more complex relationships. An example conceptual search is shown in Figure 5.3. Each circle is either a result set operation (Λ or V) or a semantic base query (?). The root operation produces the user's final result set.

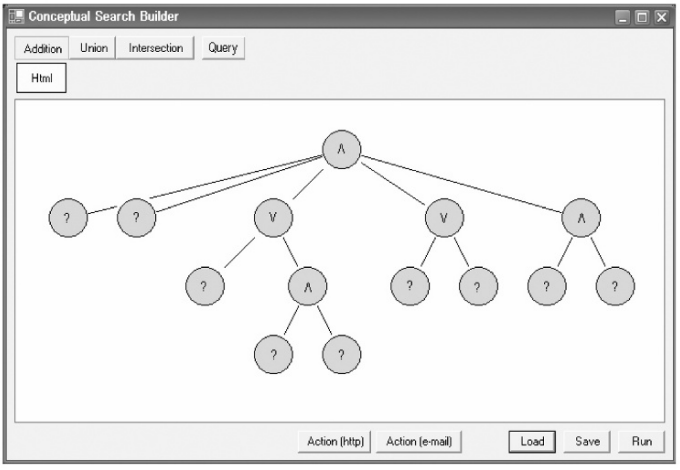


Figure 5.3: A screenshot from the conceptual search builder; each conceptual search is composed of tiered layers of result set operations

The real power of conceptual search comes from assigning values to each URI. These values can then be combined along with the URIs to form complex utility functions or formulas for defining preference. Values are assigned in three ways.

First, URIs can be assigned a constant value. For example, all airports that are within active NOTAMs may get a value of −5. Second, URIs can get a value based on how many times it appears in a result set. This can be used to count how many times the URI is related to another bound variable. For instance, a query for all airports and their C5-capable runways would give each airport a value based on how many C5-runways it has. Finally, a value can be based on the literal object of one of the URI's properties. This is highly useful for geospatial conceptual searches. An airport URI could be assigned its distance from DC or its elevation above sea level. Conceptual search thus allows a user to assign values to URIs based on characteristics of the query result set.

These values can then be altered with weights and thresholds. Each connection between query result operations can have a weight, a threshold or both. The weight is a number that is multiplied with each value in the result set to produce a new value. Then, if any value is below the threshold, it and the URI are dropped from the result set. These weights allow the normalization of values and the definition of preferences. The user needs to normalize values to provide meaningful comparisons between query results. If a user wants to add the values given to an airport by the number of runways and the distance from DC in meters, the addition may be 2 + 10,000 = 10,002. The number of runways is thus minuscule in comparison to the distance to DC. The user can use weights to increase the value of the runways (say by 4,000) to make the comparison more equal (as 2 × 4,000 + 10,000 = 18,000). Weights can also signify preference and utility functions in the same way by making one result set's values worth more than another's. For instance, a user may prefer airports in Virginia to those in Maryland and thus give a higher weight to airports in Virginia. Each result set is thus weighted and compared until the URIs reach the root of the conceptual search. The final result is the root's set of URIs sorted by their values. In the example, the result of the user's conceptual search would be the airport with the highest preference value. The user is thus able to query for a subjective concept such as preference by reducing the concept to a base set of semantic queries and expressing the relationship between them via an ontology.

5.5 Implementation

Conceptual search is revolutionary not only in allowing complex query definition but also by providing a new model for query execution, as shown in Figure 5.4. Conceptual searches are processed by a set of Web services. First, a thin client provides a graphical interface for constructing an instance of a conceptual search. This client then sends the conceptual search to the Conceptual Search Service that decomposes it into a set of component semantic queries. Each of these semantic queries can then be transmitted to a different semantic repository, which in turn may access semantic and nonsemantic data sources (such as WFS). The results are then returned to the Conceptual Search Service, which processes and returns the query result sets. Both the thin client and Conceptual Search Service are available on SemWebCentral (Dean and Kolas 2005).

A service-oriented architecture (SOA) has several advantages that support the goals of the Semantic Web. The SOA approach allows individual Web services to be added and swapped seamlessly. That is, if a new conceptual search processor is developed that still adheres to the conceptual search ontology, specific thin clients do not need to change. More important, though, is the addition of new data sources. If each data source has a corresponding OWL-S (Marin et al. 2004) representation of the service, data sources can be added with no cost. OWL-S defines three things:

what a service does, how it works and how it is used. This information is all the conceptual search needs to query a data source. The OWL-S describes which types of data the data source provides (like airports) and how to access it (like WFS). The OWL-S API (Sirin) from MINDSWAP is used to dynamically query data sources using their OWL-S descriptions. A new data source can be added or changed without affecting conceptual searches. The Conceptual Search Service will automatically incorporate the new or updated data source in subsequent queries through dynamic invocation.

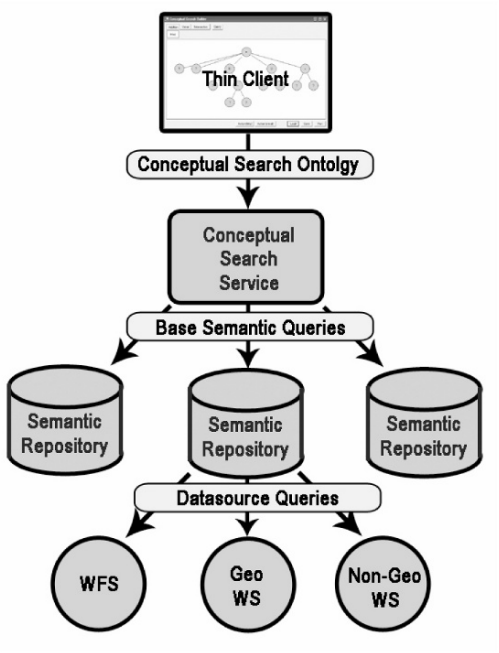


Figure 5.4: Each conceptual search can utilize a large number of services as the thin client sends the conceptual search to a service, which sends the base semantic queries to a set of semantic repositories, which can then send data-source queries to Web services

Additionally, conceptual searches can be defined once and run repeatedly. A user does not want to be tied to a workstation, continuously polling for new data. For example, airport and city location data are very static, while weather and NOTAM data may change hourly. Thus, a conceptual search can be defined once, stored in a knowledge base and periodically run for the analyst by the Conceptual Search Service. The Conceptual Search Service could also be run only when there is a noteworthy event, as described by Cole and Hornsby (Cole and Hornsby 2005). The service then alerts the user via a Web page or email when there are new results. Users can thus regularly check a Web page to see what their current preferred C5-capable airport near DC is or be alerted when their preference changes. A sample Web page output is shown in Figure 5.5, where the URI, its value and some of its properties are posted. Analysts are thus free to find new data sources and review results rather than execute queries.

The Preferred C5 Airport for Washington, DC

http://ww3.geoenterpriselab.com/DafifGF.owl#DAFIF_ARPT.6941	3.054

Properties

http://gsw.projects.semwebcentral.org/2005/03/airport/airport-ont#icaoCode	KADW
http://gsw.projects.semwebcentral.org/2005/03/airport/airport-ont#airportName	ANDREWS AFB

Figure 5.5: A sample Web page showing the results of a conceptual search; this page can be updated continuously via a standing conceptual search

Conceptual searches can be used for more than defining preferences. Any application that needs comparisons of query result sets according to a utility function would benefit from conceptual search. For instance, conceptual search can be used for hypothesis testing. A hypothesis specifies a set of supporting evidence that must be present for the hypothesis to be true. An example hypothesis is that there are new border tunnels being dug from Mexico to the United States. A conceptual search would thus be built from base semantic queries such as are there seismic disturbances near the border, are trucks hauling dirt large amounts of dirt and are sinkholes appearing from underground construction. Some of this data is more indicative of a border tunnel than others. Thus, a user may assign a larger weight to seismic disturbances, since they heavily indicate underground digging, and a lower weight to trucks hauling dirt, as those may happen for any common construction site. An analyst would therefore be alerted if there is any presence of seismic data or if there is a lot of truck movement. The user would only need to set up this conceptual search once and would automatically be alerted when there was supporting evidence of a new border tunnel.

5.6 Conclusions and Future Work

Conceptual search gives users the ability to query for abstractions relevant to them, especially in the geospatial domain. A conceptual search is based on low-level semantic queries that invoke data sources by translating a user's query into semantic and nonsemantic data-source queries. These results are then integrated via GeoSWRL to produce statements in the user ontology. Finally, conceptual search interprets the results of these queries by assigning values. Conceptual search also uses Web services and OWL-S to flexibly change data sources and permit standing queries. Conceptual search thus presents an improvement in both expressivity and execution compared to both traditional and semantic queries.

Future implementations of conceptual search may go beyond an SOA approach to incorporate emerging Semantic Web frameworks. Haystack (Quan et al. 2003), for example, builds semantic services with specific RDF processing. These services can then be run automatically to continually test hypotheses and run conceptual searches. More generally, autonomous agents (Vahidov 2005) can be used to perform each result set operation and query in a conceptual search. Each circle in Figure 5.4 would represent an agent that would receive inputs from multiple agents and feed its own output to another agent. This approach would facilitate conceptual search management by reusing specific agents and distributing the necessary resources across many systems. Any new implementation of conceptual search would require considerable work but would be able to leverage the current ontologies and services in the current SOA approach. These new implementations could also leverage current research in service composition, context matching and user profiles. Scalability and performance tests could also be run against these implementations to compare their efficiency.

Chapter 6
Location-based Web Search

Dirk Ahlers • Susanne Boll

Abstract. *In recent years, the relation of Web information to a physical location has gained much attention. However, Web content today often carries only an implicit relation to a location. In this chapter, we present a novel location-based search engine that automatically derives spatial context from unstructured Web resources and allows for location-based search: our focused crawler applies heuristics to crawl and analyze Web pages that have a high probability of carrying a spatial relation to a certain region or place; the location extractor identifies the actual location information from the pages; our indexer assigns a geo-context to the pages and makes them available for a later spatial Web search. We illustrate the usage of our spatial Web search for location-based applications that provide information not only right-in-time but also right-on-the-spot.*

6.1 Introduction

Even though the Web is said to be the information universe, this universe does not reveal where its information bits and pieces are located. Web information retrieval has long focused on recognizing textual content of Web pages and supporting keyword-based Web search. Just recently, also driven by the advent of mobile devices, the relationship of information to a physical location has gained a lot of attention on the Web. The geographic location of the user became a key element in providing relevant information "here and now". Many so-called Web 2.0 applications such as Flickr,[31] Plazes[32] or Upcoming[33] strongly relate their content to a physical location by means of manual tagging. Such manual annotation is not reasonable on the large scale of all Web pages but is not necessarily needed. Web pages such as home pages of businesses, restaurants, agencies, museums etc., but also reviews, link lists or classifieds directories, already contain location-related information. The relation to a physical location is not semantically captured but is implicitly part of the content as an address or a place name. Even though such pages represent only a fraction of all Web pages, their relation to a location is a yet unused asset for an interesting set of location-based applications.

In this chapter, we present our approach to automatically derive the spatial context from Web pages and contribute to the challenge of location-based Web information retrieval. Our technical foundations for spatial Web search comprise the methods and components for a spatial search engine – a focused crawler, a location extractor, a geocoder and a spatial indexer. The challenge is to reliably identify location-bearing pages, precisely extract the desired information and assign a geo-context to them. Even though our approach will not be able to spatially index all Web pages, it still allows for extracting geospatial context from unstructured Web resources so that a large set of Web pages can be automatically geotagged and employed in a variety of location-based applications.

The structure of this chapter is as follows. We briefly present related work in Section 6.2 before we introduce the reader to the challenges and potential of location-based Web search in Section 6.3. The architecture of our location-based Web search engine is introduced in Section 6.4, and key concepts are presented in Section 6.5. We present our demonstrator applications illustrating spatial search and discuss our experimental results in Section 6.6 before we come to a conclusion in Section 6.7.

6.2 Related Work

The related work can be grouped into three different fields: efforts to standardize the description of location information, existing location-tagged content, and commercial spatial search engines and recent research efforts for automatic extraction and indexing of location information from Web pages.

Today, several standards for description and exchange of location data on the Web exist with various powers of expression. These range from simple coordinate-oriented ones specifying latitude and longitude such as vCard (Dawson and Howes 1998), Microformats[34] or W3C Geo (W3C 2003) to more powerful formats able to express additional concepts like lines, boxes, polygons etc. such as Dublin Core Metadata,[18] the Geography Markup Language (GML)[28] or the KML format[35] of Google Earth.[2]

While older formats use simple text formats, recent formats are based on RDF or XML vocabulary, thus allowing them to be integrated into any XML document. Semantic approaches are under way to integrate spatial entities and their relations into OWL (W3C 2006b). The description of a location is typically accomplished in two ways: specification of a globally unique coordinate tuple (i.e., longitude, latitude and optional height, usually in the WGS84 frame of reference) or a named hierarchical description.

Existing geo-referenced data on the Web is mostly manually annotated or tagged. Services like geourl.org (Hansen 2006) parse specific HTML metadata specifying a coordinate enabling location-to-URL mapping; plazes.com and Placeopedia[36] are community-driven efforts to add location to content. Photo sharing sites such as Flickr[31] and Mappr[37] allow geotagging of images. Additionally, Web directories such as dmoz.org[38] organize Web links according to geographical classification. Hierarchies of places and place names are provided by so-called gazetteers, for instance (Getty Trust 2006).

Spatial search today is already provided by services such as Google Maps,[39] Yahoo! Maps[40] or MSN Live Loval.[41] Most of these rely heavily on classifieds directories and perform only little actual search to gather their points of interest (POIs), so their spatial ability is not directly coupled to the Web. For visualization on a map, addresses have to be converted to coordinates by geocoders such as the free US-geocoding service.[42]

Research towards extracting and assigning geographic meaning to plain non-tagged Web pages has gained attention in recent years. Graf et al. (2006) give a broad overview of state-of-the-art in this field. In the following, we select only those papers that are very recent and/or strongly relate to our field of work: learning geographical aspects of Web resources' contents as well as using third-party search engines for this task are covered by Ding et al. (2000). Markowetz et al. (2005) describe various challenges in identifying geographical entities as does McCurley (2001). The challenges associated with focused Web search in general are outlined in the works of Chakrabarti et al. (1999), Diligenti et al. (2000) and Tang et al. (2004). The most

recent work by Gao et al. (2006) addresses the use of geographic features for an improved multimachine crawl strategy.

Complementing existing approaches, we focus on extracting and indexing exact geographic points, in contrast to other work in the field that addresses broader geographic areas and entities.

6.3 Enabling Location-based Web Search

While most current search engines are very efficient for keyword-based queries, querying for Web pages relating to a certain location is not yet widely available. At the same time, location has a high significance for the user. A study in 2004 (Sanderson and Kohler 2004) finds that as much as 20 percent of Web queries have a geographic relation, with 15 percent directly mentioning a specific place.

The goal of a geographic search engine must be to best identify those pages that have a relationship to a geographic location, analyze and process it, and make it accessible for spatial search, which can then answer queries for relevant information at or near a certain location. The first step towards such a search engine is to find a source of spatial information. The physical infrastructure of the Internet reveals little about the geographic aspects of its contents. Estimates based on IP address or DNS entry of the servers can reveal the location of parts of network infrastructure but are seldom related to the information stored on it, especially for large hosts with thousands of domains per server. Exceptions may be dedicated Web servers for large companies; e.g., The New York Times as found by McCurley (2001) or Markowetz et al. (2005).

Therefore, a geographical search has to rely almost entirely on the contents of the information sources for information discovery; other techniques such as link graph analysis are outside the scope of this chapter. Fortunately, plenty of Web pages already contain viable location information, but not in a semantically structured way. According to "experiments with a fairly large partial Web crawl", (McCurley 2001) found that "approximately 4.5% of all Web pages contain a recognizable US zip code, 8.5% contain a recognizable phone number, and 9.5% contain at least one of these". Generally, location information can range from a brief mention of a region or a precise reference to a specific place.

Using Web information retrieval methods, we aim to analyze Web pages' unstructured contents to identify and extract geographic entities (geoparsing). If we can identify these, we are able to assign a geographic location to the Web page it was found on (geocoding). Previous work in the field of geographic information retrieval has often dealt with the extraction of regional or local coverage. Based on our research background in mobile applications and pedestrian navigation (Baldzer et al. 2004), we present an approach that aims for high-precision spatial information. We focus on the geographic entity of an individual address of a building identified by its house number.

Since the Web is a very large body of documents, it is not possible to visit and geoparse every single page. The challenge is to identify and predict those pages containing relevant geographic data within the growing and increasingly dynamic Web. We therefore developed different strategies to narrow the amount of pages we have to analyze. With our approach we create not only a geographically aware search engine, but also a truly regionalized crawling strategy. This makes it possible to create a specialized local search engine that can scan a confined region reasonably fast, missing only a few relevant pages, and so quickly create regional search coverage.

6.4 Overall Architecture of Our Spatial Search Engine

We designed and developed a search engine that enables geospatial search on the Web. We follow the general architecture of a search engine, but offer specific alterations and additions for the geographical focus. The typical components of a search engine can roughly be described as follows:

- A *crawler* discovers and downloads Web pages. The crawler takes a URL, downloads the page, analyzes the content and repeats this by following outgoing links.
- An *indexer* tokenizes the page, identifies relevant tokens and makes them accessible to search by storing them in an index.
- A *front end* allows the user to search the index with a submitted query. It handles query processing and presentation.

Our search engine is illustrated in Figure 6.1, especially highlighting the components that enable the geographically focused crawling and location-based indexing and search.

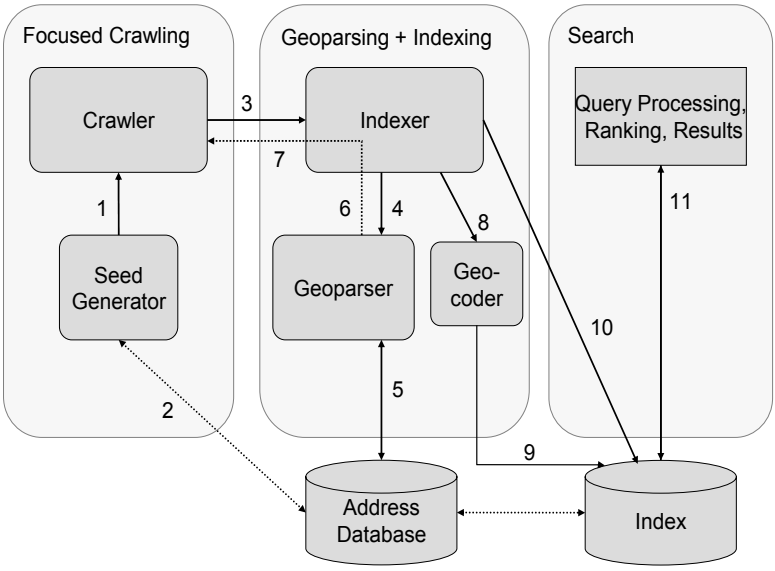


Figure 6.1: Architecture of the spatial search engine

To enable an efficient geographically focused crawling, a seed generator feeds (1) the focused crawler. These seeds are location-oriented so the crawler will start with pages that have a high probability of a relation to a physical location. For creating the seeds, an address database can be used (2). The crawler hands its downloaded pages over to the indexer (3). This relays the page to the geoparser (4), which tries to extract geographic information. This extraction process is supported also by an address database (5). Whenever relevant location-related information is found, this is indicated (6) to the indexer for later geocoding and insertion into a spatial index. At

the same time the information about a found location on a Web page is fed back (7) to the crawler to support the focused crawling. If the geoparser identifies one or multiple addresses on a page, each address is then geocoded, i.e., the hierarchical textual description is mapped to a coordinate of latitude and longitude (8) by a geocoder using a commercially available address reference table. The coordinate is then stored (9) along with the textual address and the Web page's URL. The page itself is also indexed (10). Both textual and spatial index can then be used for the known keyword-based search, now in combination with spatial search queries (11).

6.5 Central Concepts of the Geographical Search Engine

The two main concepts of our geographical search engine are geographically focused crawling and geoparsing. These two components serve to retrieve pages with geographic information and extract it.

6.5.1 Focused Crawling

While a usual crawler will simply crawl the Web breadth-first, retrieving all linked pages, a focused crawler is designed to retrieve only those pages related to a certain topic. The goal is to quickly build an index of most of the relevant pages on a topic while keeping the processing of nonrelevant pages to a minimum. The driving observation behind this idea is that links between Web pages are often set rather purposefully by their authors to link to pages containing similar information. Thus, for a given topic, strongly correlated subgraphs within the Web exist, comprising a large amount of all available Web pages on that topic.

To efficiently stay on-topic, the crawler is started with a set of pages from this subgraph (seeds). During the crawl it is steered and controlled by analyzing downloaded pages and only further processing pages that match the given topic as determined by a classifier. Observed in Diligenti et al. (2000) and Chakrabarti et al. (1999), a strong direct graph cohesion does not always exist. Highly relevant pages may be separated by nonrelevant ones. Therefore, a certain amount of nonrelevant pages, so-called bridge pages, has to be accepted to maintain a broad view of the relevant pages.

6.5.1.1 Geographically Focused Crawling

A focused crawler specializing on a spatial topic poses some special challenges: while classic topics for focused crawling ("database systems", "health", etc.) are usually strongly interlinked, this is not necessarily true for the spatial topic. Pages with an address in a certain region do not necessarily directly link to each other. Most links are rather set between similar topics and not between similar locations. Still, a sufficient number of links exists between regional pages, but with a lower density. Therefore, the specific information we are interested in is not found in dense clusters, but rather in weakly connected networks. This means the crawler has to crawl a far greater radius, spanning many bridge pages to properly cover the spatial topic and reach relevant pages.

Starting with a given region like Oldenburg that is to be crawled, our classifier determines pages to be on-topic if an address of the given region is found. To deal with bridge pages, we assign a score to pages with addresses. For each followed link, the score is restored if an address found; otherwise the score is decreased. If the score falls below a certain threshold, the current crawl branch is pruned and no

more links are extracted. Our experiments with different values for the amount of bridge pages show that their impact varies strongly depending on seeds and discovered domains but that an overall limit of bridge pages of two to five pages yields promising results.

6.5.1.2 Selection and Generation of Geo-Seeds

For geospatial search the seeds are pages that have a strong relation to the region of interest and are well linked to other on-topic pages. A well-centered seed such as the home page of a city in the targeted region or a list of local museums etc. considerably reduces the time it takes to reach other pages in that region. Two different seed types can be distinguished:

- Similar to other search engines, a first strategy to gain viable seeds is *directory-driven*. We take seeds mainly from dmoz,[38] a very comprehensive human-edited hierarchy of Web pages. The geographic hierarchy is examined, and pages located in the region of interest are retrieved. This ensures that a lot of relevant pages are already included in the seeds or are only very few links away.
- A second strategy for seed generation arises from the observation that current search engines are very good at keyword-oriented search. We shift the workload of the architecture from crawling to seed selection. We construct queries for each individual street within the desired area by using lists of all relevant cities, zip codes, and street names. The queries are sent to an already-available large index of Web pages from a major search engine and result in pages containing address parts for the region of interest. We call this keyword-query-driven approach *focused seed generation*.

Using the second strategy, the seeds themselves already contain location information. If only these seeds are analyzed and processed without further link extraction or crawling, the process can reach a coverage for a given region very fast, at the price of false dismissals that can occur for pages that are never downloaded. Additionally, it is able to use most index-backed search engines as data sources for aggregation. We call this rapid-result-oriented approach *Inverted Geo-Crawling*.

Geographically focused crawling can thus be realized using different strategies to retrieve relevant location-bearing Web pages. While the strategies cannot ensure that all retrieved pages are on-topic, they have a far higher yield than an unfocused crawl strategy.

6.5.2 Extraction of Geo-Information from Web Pages

The main challenge of geoparsing is to determine a Web page's geographic context and pinpoint an exact location. As we focus on the geographic entity of a specific address, our geoparser needs to identify zip code, city name, street name and house number. We target our approach to German addresses as our main area of interest. With reasonable effort, the geoparser can be adapted to other countries. In the following, we present our geoparser's methods for disambiguation and validation of address information from Web pages.

To reliably extract an address from a Web page, its individual parts have to be identified. The parts are not necessarily unambiguous; for disambiguation, individual parts have to be considered in relation to each other to ascertain a full address using various heuristics described in this section. Many geographic applications use a gazetteer for reconciliation with existing knowledge about geographical entities

such as places, regions, countries, cities. For improved accuracy, we take this strategy one step further and use a full database of address-related information, which contains zip codes for every possible city and also every city-zip combination for each street.

McCurley (2001) describes certain ambiguities particular to geographic entities that arise even with assisted search. *Geo/non-geo ambiguity* refers to the use of terms to name a place as well as a different concept, e.g., the German word "leer" means "empty", but is also a small town in East Frisia; "Münster" means a minster, but is also the name of several cities. The second example also illustrates *geo/geo-ambiguity* where different places share one name. These are cases where traditional keyword-based search most likely fails, since a single named entity can only be reliably located with additional location information. The appearance of, e.g., zip code and city name near each other generates a stronger geographic hint to a certain city. The same applies to street names. The zip code in itself is ambiguous as well: a German five-digit zip code could also be a product or phone number, a price, etc.

Starting with the zip code as the main supporting term, we initiate a coarse-to-fine term disambiguation. We draw upon an address database for reliable identification of terms by searching for city names in the close surrounding of a found zip code on a page. Once zip code and city are identified, we extend the disambiguation towards the street level by searching for street names for the city-zip pair. Finally, if a house number is found, the geoparser treats the address as valid and geocodes it.

City and street names on Web pages do not always match the names we have provided in the database. We therefore employ normalization and stemming methods to be immune against variations. For normalization, name additions or city districts are given a lower relevance to also match cities where this was omitted. Spelling variations are allowed to correct possible typos. For the detection of street names, they are subjected to stemming algorithms to reduce the street name designations (e.g., "Strasse" – street; "Allee" – avenue, etc.) to a single token as these are often abbreviated in various ways. Separation of name parts such as hyphenation, spaces, written as one word or mixture of this is identified. Again, spelling variations are considered.

This extraction method has the advantage that the extraction process is tied strongly to existing knowledge and only valid addresses with parts known to be correct are extracted. Some of our lessons learned while implementing the geoparser were used to improve the keyword-query-based seed generation described earlier.

6.6 Experimentation and Demonstration

Our proposed methods and strategies for crawling and geoparsing were implemented as prototypes to gain experimental results and support our design decisions. We discuss our results for different crawling and seed selection strategies and present the results for the location assessment.

6.6.1 Evaluation of Geographically Focused Crawling

Based on the two proposed seed selections, we ran several tests to show the validity of our approach. We present the results grouped by the method of seed selection as the seeds constitute the main input for the crawling strategies.

6.6.1.1 Results of Query-based Seed Selection

We ran a crawl with query-based seeds in the regions of Oldenburg and Rügen. Oldenburg is an urban city, and Rügen, Germany's largest island in the Baltic Sea, is a rural area. These were chosen to assess influences of the crawled region from the results. We generated queries from both regions' street and city data to query a search engine and let both tests run independently. For our tests, we utilized the Google API,[43] which we chose due to the very large index size, a resulting high number of results and its public interface for queries. In this test, we only processed the retrieved pages with our geoparser; the crawler did not follow any further links (inverse geocrawling). Random manual sampling of the results was done to check the results for correct location assessment.

It took about four days to retrieve and analyze the data for the 1,379 streets of Oldenburg. The geocrawling resulted in about 24,000 addresses on 23,000 distinct pages. This means that some of the retrieved pages were directories with multiple addresses. Most of the retrieved pages contained only a single address. For 240 streets, no pages with an address were found. On average, each street was present in 17 addresses.

The results for Rügen were similar; here we started with 1,074 streets and found about 17,200 addresses on 21,100 pages in a little under four days. We found fewer directory pages, but some sites that featured the same address on each page. We found that 356 streets had no address associated. An average of 16 addresses per street was found.

Generally, the results for both regions are quite similar in structure but differ in quantity. We found a full address on 25 percent of all downloaded pages. About 75 percent of all raw results were discarded by our parser; thereof 90 percent because of missing cohesion (it is not currently possible to search for term nearness with the Google API, so a lot of pages were obtained with the search terms scattered about the page without forming a contiguous address) or because the address could not be found on the page at all. The remaining 10 percent were dropped because the page was no longer available, no house number for an address could be found, or other issues. Fewer than 10 percent of pages with otherwise correct addresses lacked a house number.

We could prove that fast coverage of a region is possible with the inverse geocrawling approach of query-based seed generation and that it is a reliable way to quickly build an index of a confined region, uncovering relevant pages in a short frame of time. We also showed that our approach of only allowing full addresses including house number is valid to discover high-quality addresses.

6.6.1.2 Results of Directory-based Seed Selection

We tested our approach of geographically focused crawling with directory-based seeds from dmoz[38] for the large region of northern Germany. As opposed to the query-based seed selection, the seeds were mainly hub pages containing only very few addresses themselves. We therefore fed these seeds into a crawler with a generally unlimited crawl depth and high bridge page number, but some filters activated: a maximum number of documents per domain was in effect, as well as filters to exclude non-German domains, unpromising contents such as galleries or forums, etc. For a crawl of two weeks, the results show that of about 44,000 domains crawled, 20,500 contained at least one full address in the region with the address count at 3.8 per domain. For this domain-oriented analysis, we only counted unique addresses

and complete domains. For a rather broad crawl, we feel that finding addresses on about half of all visited domains is a good result. In a second step, we set up smaller crawls of Oldenburg with a small seed to directly examine the effect of different bridge page parameters. We found that when comparing an unfocused to a focused crawl for this region, we can retrieve up to 10 times as many addresses with a focused crawl in a given time.

Our geographically focused crawling was dependent on several factors: for small regions, the approach identified much more Web pages with addresses in a given time than a crawler without this focus. For larger geographic regions, however, the number of bridge pages had to be much higher to find enough addresses; even then, the ratio was smaller than for confined regions. This indicates that the crawl tree still dilutes quickly at some point, and gathers increasingly more pages that are off-topic. We currently work on larger crawls to give better estimates on the ratio of Web pages containing addresses on a larger scale.

6.6.2 Quality of Geo-Information Extraction

Some numbers on distribution and structure of address-bearing Web pages were already mentioned in relation to our crawls where the geoparser was used to extract addresses. Discussing quality measures for the address extraction itself, precision was found to be very high. Of the identified addresses, random sampling reveals almost no errors. The presented methods leave only very little room for misidentification and incorrect addresses are not recognized by our parser, so this is to be expected. Recall is difficult to measure as we cannot make reliable assumptions on the number of all relevant documents. We are aware, however, that there are certain omissions due to addresses that we cannot currently find. Our methods assume a strong cohesion between address parts by using maximum distances that are exceeded by certain pages, often by elaborate table structures that dilute the relationships between address parts. A typo in a zip code can invalidate a whole address. Finally, multiple typos or unusual abbreviations cannot always be matched. We already tuned the heuristics so that weakening them more would lead to erroneous addresses. Due to these effects, we estimate an omission rate of 5 to 10 percent.

We found that our database-backed approach still outperforms simple address matching as could be done by a general matching such as “street term + number + zip + city term”. For the city of Oldenburg alone, we found that 118 of 1,725 streets did not match any usual street name pattern, which was already extended to include local designations like “Kamp” (field). Some of these are decidedly un-street-like such as “Ellenbogen” (elbow), “Ewigkeit” (eternity), “Vogelstange” (bird perch) and might even be prone to misrecognition with additional supporting terms. Discovering these with a list of known names leads to better location coverage.

6.6.3 Applications

The spatially indexed Web pages gathered by our search engine enable some interesting applications. Generally, we install a spatial layer on top of the existing Web. This layer captures the semantic location information of the pages. Pages can be located at certain coordinates, and a search can be restricted to a desired area. We implemented two applications that illustrate the potential of a spatial search engine in two different application domains: first, a location-based search for locating Web pages on a map; second, a search engine for the enrichment of directory data of, e.g., yellow pages.

6.6.3.1 Localized Web Search

Our first prototype exclusively used query-based seeds for page discovery and directly processed these with our geoparser. It was built as a feasibility study of geo-extraction and -referencing. Using a commercially available geocoder from a related project, extracted addresses were mapped to geographic coordinates.

Figure 6.2: Local search results for Oldenburg

For our main demonstration we chose the city of Oldenburg. With our search engine, users can search the content of all pages with an address in this city. The screenshot in 2 shows the prototypical result page of a search for "pizza". The search box at the top of the page repeats the query; below is the map with the results as numbered icons. Only pages with an address in Oldenburg matching the keyword search are displayed. The zoom level of the map adjusts according to the distribution of the result. Below the map the results are listed for each icon number along with the page's title, URL and the identified address. It is immediately clear where the different pages and thus pizza services are located. Our prototype thus demonstrates a keyword-based Web search with spatial awareness.

6.6.3.2 Spatial Search for the Enrichment of Directory Data

While many local search applications rely on classifieds directories as data sources, the search engine for this project works the other way around. For our project partner, a major provider of yellow pages, we built a search engine that can automatically enrich existing directory entries and also discover promising new candidates. This will act as an additional source of data for the survey department, which until

now had to rely completely on manually gathered data. By automatically retrieving business Web sites, existing yellow page entries can be enhanced and sanitized and prospective new customers can be uncovered. The screenshot in Figure 6.3 shows the query interface and a detailed result of a search for the location of our institute.

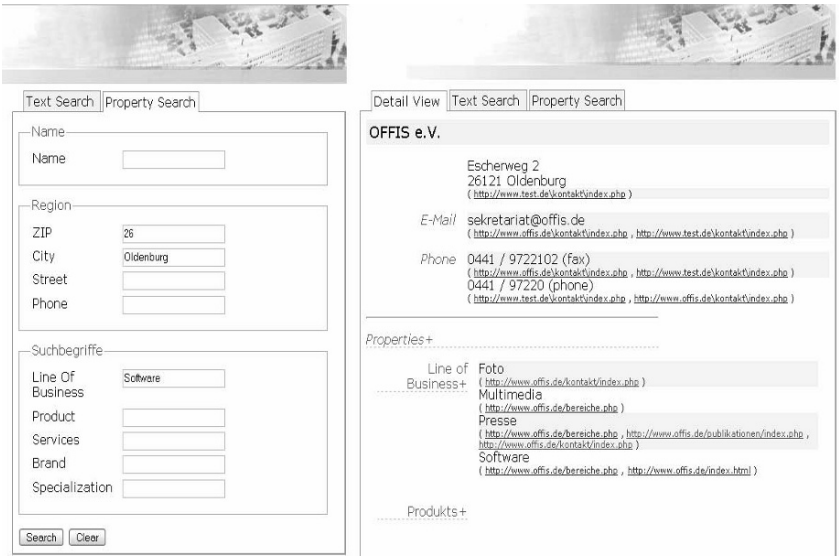


Figure 6.3: Directory data query form and result

We based this search engine on a standalone crawler for resource discovery. We implemented a general focused crawling to capture a large part of northern Germany with seeds selected from dmoz.[38] The geoparser is the initial component, which is extended with other parsers to derive additional information from the Web pages: we built an entity name extractor that can derive the person, company or organization that is referenced by an address. We also extract business specifications, such as commercial register entries, phone numbers, line of business, products, by abstracting from individual Web pages.

This automatic data acquisition is now used commercially to greatly improve quality and search time, since a much better and faster overview of companies is gained.

6.7 Conclusions

In recent years, location has gained much attention in a Web context. Rather than expecting mass Web content to be manually annotated with its location, the implicit location relation already present on plenty of Web pages that lay unused now forms a valuable asset for a range of commercially relevant applications.

In our approach, we developed central concepts and components for a geographic search engine: a focused spatial crawler and a geoparser that identify precise localized information and input it to a spatial index that complements the keyword-based index of today's search engines. We now have the ability to perform a keyword- and location-based search of common Web pages by filtering the index according to geographic properties. Our experimentation gives interesting results that

we will also refine in our future work. The data sets crawled for Oldenburg and Rügen gave an interesting insight into performance of the algorithms and precision of the found and spatially indexed Web sites in these regions.

We applied our spatial search engine in two different application domains. A localized search (profiting from our regionalized approach) can offer a spatial search on the Web and not only on previously selected and annotated pages. Based on keywords, the results show the spatial relationship of the found Web pages. The second application domain might not be so obvious on first sight. The localized search results are used for extending and sanitizing existing yellow page databases. This approach has been very successfully carried out with a research cooperation partner. Future work will mainly concern broadening the scope, adding more semantics and building stronger relationships and rankings.

Acknowledgements. We thank our colleagues Jörg Baldzer and Norbert Rump for their continuous support and collaboration on this project as well as our student Dorothea Eggers for her valuable work. Part of this research has been carried out as a subproject of the Niccimon project (Scheibner et al. 2006) and was supported by the State of Lower Saxony, Germany.

Chapter 7
Ubiquitous Browsing of the World

Gabriella Castelli • Alberto Rosi • Marco Mamei • Franco Zambonelli

Abstract. *Pervasive computing technologies, together with the increasing participation of the Web community in feeding geo-located information within tools such as Google Earth, will soon make available a huge amount of real-time information about the physical world and its processes. This opens up the possibility of exploiting all such information for the ubiquitous provisioning of context-aware services for "browsing the world" around us. However, for this to occur, proper general-purpose data models and software infrastructures must be developed. In this chapter, we propose a simple, yet effective model for the representation of heterogeneous contextual information and the design and implementation of a general user-centric infrastructure for ubiquitous browsing of the world. The presentations of some exemplar services we have implemented over it and have made available to users via Google Earth interfacing complete the chapter.*

7.1 Introduction

The increasing diffusion of embedded pervasive computing technologies such as RFID tags (Want 2006) and sensor networks (Chong and Kumar 2003) will soon make available a huge amount of real-time information about the physical world and its processes. In parallel, the success of participatory Web-based geospatial tools (e.g., Google Earth) is leading to the mass production of geo-located information coming from diverse communities and related to a variety of facts and events situated in the world (Butler 2006). These trends contribute to accumulate information that can be used to build real-time and historical models on a number of facts and processes happening in the world around us. In other words, they open up the possibility of "browsing the world" (Castelli et al. 2006), i.e., integrating information coming directly from surrounding pervasive devices as well from the Web and exploiting it to effectively support, in a ubiquitous and context-aware way, any activity related to understanding the world and interacting with it.

In the above rapidly developing scenario, it is desirable to avoid the proliferation of ad hoc solutions and systems. Rather, efforts should be devoted to identify solutions facilitating the engineered design and deployment of general-purpose "browsing the world" services. Specifically, we think it is of fundamental importance to identify general-purpose – expressive yet simple-to-be-manipulated – data models, together with proper software infrastructures to organize data and provide access to it by services.

Starting from the above considerations, the first contribution of this chapter is to propose a simple data model to represent "facts" about the physical world, for the use of both users' querying activities and context-aware browsing the world services. The model, which we call "W4", is based on the consideration that most information about the world (whether coming from embedded sensors, tags or communities) can be simply represented in terms of four "W"s – *Who, What, Where, When* –

and that such a representation enables for very expressive and flexible context-aware data usages.

As a second contribution, we describe the design and implementation of a general middleware infrastructure for browsing the world, facilitating the development of and supporting the activities of general-purpose context-aware services. The infrastructure supports PDA and laptop access to information coming from both pervasive devices and the Web, provides for representation and organization of data in W4 terms, makes available a Java interface for users' queries and for services access to such data and is integrated with Google Earth and Google Maps for effective user interfacing.

This chapter is organized as follows. Section 7.2 analyzes the browsing of the world scenario and identifies the key research challenges it implies. Section 7.3 presents the W4 data model, while Section 7.4 details the implemented middleware infrastructure. Section 7.5 presents some exemplar services we have implemented on top of our infrastructure. Section 7.6 discusses related work and concludes.

7.2 Towards Ubiquitous Browsing of the World

Our everyday environments will be soon densely populated by a variety of embedded devices such as RFID tags (Want 2006) and sensor networks (Chong and Kumar 2003). Users in an environment will be able, via wireless interfaces mounted on some wearable computing devices (e.g., a PDA or a smart phone), to directly access devices in their proximities and acquire information about the characteristics of the physical world around them and of its processes, much more than their normal five senses could allow. For instance, one could get information about nearby objects by reading RFID tags attached to them (which may contain descriptive information about such objects), or one could read any kind of environmental parameters that are being sensed by wireless sensors around them.

Other than directly accessing surrounding devices, users will also be able – in most cases – to access them via the Web, exploiting some wireless communication technologies. This will enable them to dynamically retrieve a lot of additional information related to the surrounding world. Other than accessing "traditional" Web information (e.g., HTML pages and Web services), this may include geo-located information concerning specific geographical areas, general facts and annotations about them, as they can be continuously provided via collaborative Web 2.0 technologies by the Web community (Espinoza et al. 2001; Teranishi et al. 2006). Let's not forget the possibility of accessing historical sensor and tag information, as well as information generated by sensors and embedded devices located far away in the world.

Users themselves can enter the above picture by deciding to unveil – totally or to some limited extent – their presence in an environment, e.g., by making their identity, location and/or activities somehow perceivable. This can occur by dynamically uploading such information on the Web or by making it available to other users via ad hoc connections, or even by uploading it into surrounding pervasive devices (pervasive devices such as RFID tags can indeed act as a sort of distributed memory infrastructure (Mamei et al. 2006). It is worth outlining that determining the location of a specific user will be increasingly easy: even for users that do not carry a GPS, their location will be easily inferred by the patterns of access to the surrounding pervasive devices (e.g., access to an RFID tag with a known location by a user automatically localizes the user as being close to that device) (Satoh 2005).

The concept of *browsing the world*, in general, considers the possibility of navigating in an information space that – by properly merging and integrating information coming from both pervasive devices and the Web – can represent a detailed model of the world, comprising both present and historic fine-grained geo-located data about its entities, its processes and its social life. In context-aware user-centric terms, which are the ones of more interest here, the concept of browsing the world implies the following possibilities: (i) for users anywhere in an environment to access and navigate in a flexible and location-dependent way meaningful information about the surrounding physical world, and (ii) for software services to access and manipulate such information in order to autonomously adapt their behavior to users' need and their current context of use.

We are perfectly aware that ubiquitous and context-aware browsing of the world is already becoming a reality. Indeed, tools and services belonging to this category already exist (e.g., Google Earth provides the possibility of being integrated with a GPS so as to enable location-dependent navigation), and new ones appear every day (Castelli et al. 2006). However, beside special-purpose and ad hoc implementations, for browsing of the world activities and services to become a usable common practice, and for the development of browsing of the world services to become a sound engineering discipline, several challenges remain to be addressed. In particular,

- It is fundamental to identify a *general-purpose uniform data model* to represent information about the world. The model should enable representation of a variety of facts about the world, generated by a variety of heterogeneous sources (from embedded sensors to Web communities), and should be easy to be manipulated by software applications. Spatial information is clearly important, but it is not enough. Temporal information is required, as well as information describing the activities taking place in the world. Also, the model should inherently consider the existence of incomplete and/or limited accuracy information.
- It is important to define a *general middleware infrastructure*, based on the above data model, and supporting the execution of general-purpose, context-aware browsing the world services. The infrastructure should be general-purpose, autonomic and adaptable. Relying on this infrastructure, the activities of browsing the world should not be compromised because of, say, the temporal unavailability of an Internet connection, or of the GPS or of an RFID reader. Strictly related, the services running on the infrastructure should not mandate the availability of specific information but should exploit whatever information is available on a best-effort basis.

The goal of our work, as preliminary and incomplete as it may be, is to face the above challenges by defining both a simple yet effective model for world data and a general software infrastructure to support the design and execution of robust browsing the world services.

7.3 The W4 Model

A data model for expressing facts about the world should be able to uniformly deal with information coming from heterogeneous sources, should enable ease of querying and processing and should account for adaptation to context and incomplete information. Also, it should somewhat support complex semantic querying over large data sets. Along these directions, our proposal considers that diverse data about world facts can be expressed by means of a simple, yet expressive four-fields

tuple *(Who, What, Where, When)*: "someone or something (*Who*) does/did some activity (*What*) in a certain place (*Where*) at a specific time (*When*)". We also call such W4 tuples *knowledge atoms,* as they are atomic units of factual knowledge. In general, knowledge atoms may be created by proper software agents associated to data sources as diverse as embedded devices, cameras, users or Web sites and are stored in suitable shared data spaces. Users and services, from everywhere, can retrieve knowledge atoms via simple pattern-matching query mechanisms (which also support context-aware queries and incomplete information) to interact with the world and to enforce adaptive context-aware functionalities.

7.3.1 Data Representation

The four-fields (*Who*, *What*, *Where*, *When*) that constitute our data model describe different aspects of a fact.

The *Who* field associates a subject to a fact. *Who* may represent a human person (e.g., a username) or an unanimated part of the context acting as a data source (e.g., the ID of an RFID tag). The *Who* field is represented by a type-value string with an associated namespace that defines the "type" of the entity that is represented. For example, valid entries for this field are "person:Gabriella", "tag:tag#567" (for this and the other field, the choice of adopting a type-value string representation rather than an XML representation has been driven by simplicity of prototyping; the two representations are fully interchangeable and do not affect, per se, the model).

The *What* field describes the activity performed by the subject. This information can either come directly from the data source (e.g., a sensor is reading a temperature value) or be inferred from other context parameters (e.g., an accelerometer on a PDA can reveal that the user is running), or it can be explicitly supplied by the user. This field is represented as a string containing a predicate-complement statement. For example, valid entries for the *What* field are "read:book", "work:pervasive computing group", "read:temperature=23".

The *Where* field associates a location to the fact. In our model the location may be a physical point represented by its coordinates (longitude, latitude), a geographic region (we currently adopt the PostGIS language to describe such a region), or a logical place. Logical places, like "campus" or "bank", are typically described by means of another W4 tuple in which the logical place is the *Who* and the *Where* associates a specific physical location to the corresponding logical place. In addition, context-dependent spatial expressions like "here" or "within:300m" can be used for context-aware querying, as described in the remainder of this section.

The *When* field associates a time or a time range to a fact. This may be an exact time/time range (e.g., "2006/07/19:09.00am – 2006/07/19:10.00am") or a concise description (e.g., 9:28am). For example, 9:28am = 2006/07/19:9:28am ± 5min. Also in this case, context-dependent expressions can be defined (e.g., "now", "today", "yesterday", "before") and used for context-dependent querying.

7.3.2 Data Generation

The W4 model relies on the reasonable assumption that software drivers (or, more generally, software agents) are associated with data sources and are in charge of creating W4 tuples and inserting them in some sort of shared data spaces. While any data source in the end must be associated with some software to gather and store data items, W4 agents have the additional goal of collecting all the necessary information to produce a W4 tuple that is as accurate and complete as possible. This oc-

curs by sensing and inferring information from all the devices and sources available (e.g., RFID tags, GPS devices, Web services) and by combining them in a W4 tuple. Let us make two simple examples to clarify this concept: Gabriella is walking in the campus park. An agent running on her GPS-equipped PDA can periodically (e.g., every *n* seconds) create the following tuple:

Who: user:Gabriella
What: walk:4kmh
Where: lonY, latX
When: 2006/10/17:10.53am

There, the *Who* is entered implicitly by the user at the login, *What* and *Where* can be derived by the GPS (e.g., the speed of Gabriella as measured by the GPS can be used to deduce that she is walking) and *When* can be provided both by the PDA or by the GPS. Let us now assume that Gabriella's PDA is connected with an RFID tag reader. A specific RFID agent controls the reader and handles the event of "tag recognition" whenever a tag enters in the reading range. In this case, either the tag contains its own *Who* and *What* description in its limited memory or the tag ID can be resolved in a database (mapping tag IDs into the associated *Who-What* descriptions) that the agent may access to fill in the W4 fields. Otherwise, the *Who* reduces to the tag ID (which enables access to the database later) and the *What* is left empty. As in the previous example, the *Where* and *When* can be read from the user's GPS. The resulting tuple is as follows:

Who: tag:#456
What: -
Where: lonY, latX
When: 2006/10/17:10.59am

A similar RFID atom resulting from the agent having accessed the database and filled in the fields is shown in Figure 7.3.

7.3.3 Interface and Context-Aware Queries

Knowledge atoms are stored as W4 tuples in a shared data space (or in multiple data spaces). To interact with them, we take inspiration from tuple-space approaches (Ahuja et al. 1986) and define the following API:

```
void inject(KnowledgeAtom a);
KnowledgeAtom[] read(KnowledgeAtom a);
```

The *inject* operation is trivial: an agent accesses the shared data space to store a W4 tuple there. The *read* operation is used to retrieve tuples from the data space via querying. A query is represented in its turn as a W4 tuple with some unspecified or only partly specified values (i.e., a template tuple). Upon invocation, the read operation triggers a pattern-matching procedure between the template and the W4 tuples that already populate the data space. A vector of all matching tuples – i.e., those for which all the defined fields match those provided in the template – is returned as the result of the query. In any case, pattern-matching operations work rather differently from the traditional tuple-space model. In fact, our proposal can rely on the W4

structure to enforce expressive context-aware pattern-matching operations, which may exploit differentiated pattern-matching mechanisms for the various fields. Current mechanisms work as follows:

- *Who and What*. Pattern-matching operations in these two fields are based on string-based regular expressions. For example, "user:*" will match any user.
- *Where*. Pattern matching in this field involves spatial operations inspired by PostGIS operations. Basically, the template defines a bounding box (e.g., "circle,center(lonY,latX),radius:500m"), and everything within the bounding box matches the template. All tuples with a *Where* field within the circle will match this field of the template. Contextual places such as "within:300m" can be specified in the template and are translated into actual spatial regions – based on the current location from which the query is performed – before going through the pattern matching.
- *When.* In this case, the template defines a time interval. Everything that happened within that interval matches the template. Concise time descriptions as well as contextual ones (e.g., "now" or "before") are converted into actual time intervals before pattern matching.

The following two examples illustrate the query process. Gabriella is walking on campus and wants to know if colleagues are near. She will ask via a *read* operation:

Who: user:*
What: works:pervasive computing group
Where: circle,center(lonY,latX),radius:500m
When: now

Analogously, Gabriella can ask if some of her colleagues have gone to work in the morning:

Who: user:*
What: works:pervasive computing group
Where: office
When: 2006/07/19:09.00am – 2006/07/19:10.00am

It is important to emphasize that the returned answers do not have to be "complete" W4 tuples. The pattern-matching mechanism also allows matches between incomplete pieces of information. Thus, following this approach, applications are based on components entering complete and incomplete context information and getting in response refined (but possibly still incomplete) information.

7.3.4 Towards Semantic Knowledge Networks

Our current model is somewhat limited by the lack of a reference fully fledged ontology that could add semantic relationships to the concepts in the W-fields. For *Who* and *What*, we adopted a sort of informal ontology for the sake of experimentation. For *Where* and *When,* we adopted standard spatial and temporal representations, and we are at the first stage in experimenting with logical concepts of space and time. Needless to say, we plan to deeply investigate such aspects, also borrowing ideas from the existing work in the Semantic Web research.

Strictly related, we intend to enrich the W4 data model to enforce logical relationships between W4 tuples, so as to enable users and agents to query for complex facts other than for isolated knowledge atoms. The ideas are to (i) integrate in W4 data spaces sorts of "information agents" able to identify semantic relationships between tuples and link them accordingly, (ii) enable agents to query the tuple space on the basis of network relationships between tuples, other than on their individual content. With this regard, our experience so far concerns the issue of logical localization within buildings. While the *Where* of a tuple can be a logical place (as stated in Section 7.3.1), such as a room in a building, the rooms themselves can be described by tuples having the name of the room in the *Who* field and its physical bounding box in the *Where* field and/or the reference to a building map. By properly relating these tuples, it is possible for services to seamlessly switch from physical localization to logical building-map localization depending on needs and on the kinds of localization devices available.

7.4 The Infrastructure for "Browsing of the World"

To actually experiment with ubiquitous browsing of the world and with the W4 model, we have prototyped a W4-based middleware infrastructure.

7.4.1 Architectural Overview

A general infrastructure to enable human-centric browsing of the world must include services for data acquisition, data integration and data visualization. The architecture we have implemented is organized as follows (see also Figure 7.1):

- Putting humans at the center, our architecture considers users with portable computing devices (i.e., laptops or PDAs), integrating localization devices (i.e., GPS), devices to acquire information from the physical world (i.e., RFID readers and sensors) and means to connect to the Internet (i.e., WiFi and/or UMTS connections).
- Contextual information about the world, including user data, data coming from pervasive devices and more generally data available on the Web, is represented by means of W4 tuples and stored by a local tuple space to be later accessed by application agents.
- A number of additional Web-accessible tuple spaces can also be used to store and retrieve W4 information. Each space could host information related to either a limited geographic area (e.g., the campus tuple space) or a specific topic (as happens today for Google Earth – Google Maps mash-ups).
- An RFID reader (in the form of a wearable glove) connected to the laptop or to the PDA via a serial interface can be used to collect information from RFID tags dispersed in the environment. This information, enriched with the physical location where it has been collected (provided by the GPS device), is stored in the local tuple space.
- Data coming from sensor network nodes (Crossbow MICAz) can be directly accessed and stored in the local tuple space, converted in the W4 format and typically enriched with the physical location of the actual sensors. Alternatively, sensor data could be collected by a base station and sent directly to a Web-accessible tuple space.

- Specific services can be realized by means of application agents (i.e., autonomous software components) running locally on the user's portable device and accessing, via the W4 model, both the local and Web-accessible tuple spaces. Also, application agents can interface with a local GUI client (Google Earth or Google Maps) to turn data into a user-centric perspective.
- As needed, agents can dynamically connect to the Web to retrieve additional information to integrate with that coming from the W4 tuple spaces.

The whole system has been realized using the Java language. The RFID reader and the sensors are accessed via JNI and sockets, respectively. Web-accessible tuple spaces have been implemented through a PostgreSQL database with spatial extensions, while the local tuple space is implemented by a Java Vector. User interface is provided by Google Earth for laptops and by Google Maps for PDAs.

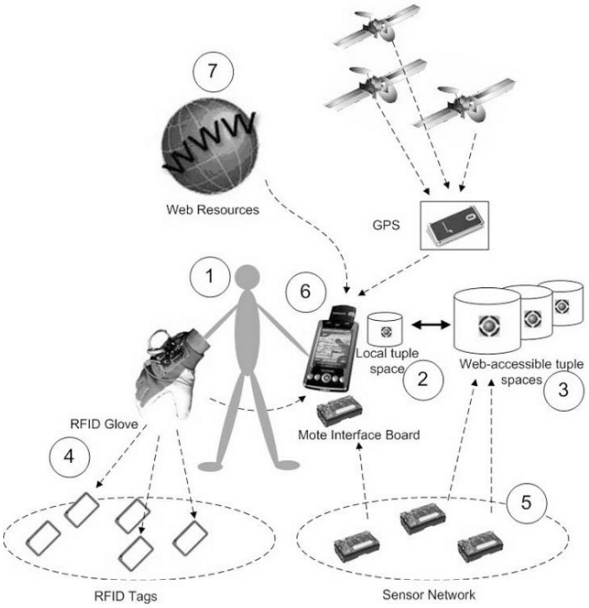


Figure 7.1: Architecture of the browsing of the world infrastructure

7.4.2 W4 Tuple Spaces

All the information coming from the embedded devices (GPS, RFID and wireless sensors) is translated by special-purpose agents into W4 tuples and stored in a local tuple space (examples of these W4 tuples can be found in Sections 7.3.2 and 7.3.3). Application agents access this space to retrieve W4 information supporting their contextual activities. Thus, application agents are completely decoupled from low-level embedded devices, and so they access and deal with contextual information only in terms of W4 tuples. In addition, the availability of a local tuple space allows the system to work also in absence of a network connection and allows minimizing the generated data traffic (and its associated costs). To have it run on simple mobile devices, the local tuple space has been implemented as a simple Java Vector accessible by agents through the W4 API interface.

The infrastructure may also involve a number of Web-accessible tuple spaces enabling more global queries. In general, an application agent performing a query accesses the local tuple space and/or may refer to a limited number of remote spaces. Our current implementation of a remote tuple space consists of a Tomcat Web server giving access to a PostgreSQL database that stores the W4 tuples. We realized JSP and Servlets implementing the W4 interface. Our PostgreSQL database is based on a single table consisting of the four Ws fields. Thus, it actually resembles an unstructured bag of W4 tuples. A general problem for our infrastructure is to identify strategies for evaluating which information to send to some global tuple space and which to keep local. Such decisions depend on many factors, such as privacy issues (e.g., a user may not be comfortable constantly sending his GPS location on the Web) and scalability reasons. In our current implementation, where the user base is limited and scalability issues are not compelling, we simply upload the local knowledge atoms to a global server whenever the wireless network is available.

7.4.3 The W4 Query Engine

The W4 query engine is the component that is in charge of managing the W4 queries, translating logical values (e.g., *When* = *now*) into actual ones (e.g., *When* = *2006/07/19:9:28am* ± *5min*) and performing pattern-matching operations. We developed two implementations of the query engine.

- The query engine running on the local tuple space has been developed in Java. Basically, it scans the local vector of W4 tuples and uses string parsing methods and simple algorithms (to handle *Where* and *When* clauses) for pattern matching.
- The query engine running on a Web-accessible tuple space dynamically translates W4 queries in SQL to execute them on the PostgreSQL database. In this implementation, query pattern matching is supported either natively by SQL or by the PostGIS spatial extensions.

Both these engines allow users to deal with spatial queries in terms of actual geographic areas (expressed in term of longitudes and latitudes) and logical places (e.g., rooms). As stated in Subsection 7.3.4, the integration of more complex semantic and network-based queries is in progress.

7.4.4 The User Interface

We developed a flexible graphical subsystem that can be easily employed on both laptops and PDAs. In particular, our GUI interfaces with both Google Earth and Google Maps to display retrieved context information as placemarks in a specific geographical area (see Figures 7.2 and 7.3). The graphical subsystem relies on the Keyhole Markup Language (KML), fully supported by Google Earth (at the moment only available for desktop and laptop computers) and at least partially supported by Google Maps and Google Maps for Mobile (also accessible by PDAs and smart phones). Simply, our graphical subsystem translates proper W4 tuples into a corresponding KML file and dynamically provides the file to Google Earth/Google Map.

It is worth noticing that the KML language also allows for specification of the user viewpoint on the map. This naturally supports context awareness, in that an agent could decide to center the map where relevant information is located. In particular, by centering the map on the user, the agent can provide a user-centric representation of the world, where the user can literally *see* nearby resources.

7.5 Application Examples

We developed some simple application services to highlight the effectiveness of our approach. In all these examples, we implemented a software agent that (i) receives either static or dynamic queries from the user, (ii) accesses a remote tuple space to retrieve suitable W4 information, (iii) creates a KML-formatted answer, and displays it either in Google Earth (for laptops) or in Google Maps (for PDAs).

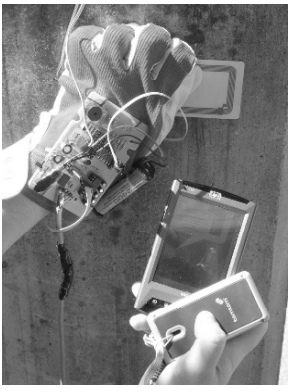
We emphasize that our goal here is not to present brand-new innovative services (services similar to the ones we present can be found elsewhere) but rather to show that the W4 model and its infrastructure can support a variety of application in a uniform and intuitive way.

7.5.1 The Journey Map

Our first application has the goal of providing real-time and historical information to a user equipped with a GPS device and an RFID reader. In particular, we focused on the scenario in which a tourist wants to automatically build and maintain a diary of his journey.

First, the proposed service allows the user to keep track of all his movements and have them displayed on the map of the visited places. Simply put, application agents can periodically (e.g., each minute) store in the tuple space the information about where the user is at the current time. Later on, the user can simply retrieve the history of his journey by having the agent query the tuple space. Eventually, application agents can interpret the retrieved information and render it as needed, e.g., as a KML-ploryline in Google Earth showing the path the user walked.

Second, the support for RFID allows tourists to access tourist information stored in RFID tags attached to monuments and art pieces. The user can implicitly or explicitly (Figure 7.2a) read tags around him. The application agent can monitor the local tuple space for the presence of W4 tuples representing nearby RFID tags. Eventually, it can display the associated information as a Google Earth placemark (Figure 7.2b). Also, if the information in the tag reduces to the tag ID, it is possible for the agent to query a Web-accessible tuple space to integrate such information.

(a) 

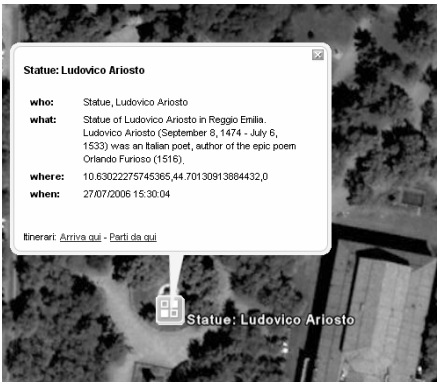
(b) 

Figure 7.2: (a) The RFID-reader embedded in a glove allows tagged objects to be identified. (b) The RFID tag becomes a placemark with integrated Web-retrieved information in Google Earth

Figure 7.3: Map showing users real-time locations with neighboring university facilities

It is worth noticing that data coming from sensor network could be accessed via similar W4 queries. The application can indeed be easily configured so as to retrieve and display environmental data coming from sensors nearby.

7.5.2 The People Map

A user equipped with a GPS device can decide to share her location with other users, and, analogously, she may wish to be aware of the location of others users. For example, a group of friends can share their actual GPS locations (represented as W4 tuples) with each other. This can happen either by uploading knowledge atoms to a Web-accessible repository or by exchanging them in an ad hoc way and storing them in the local tuple space only. Either way, collected W4 tuples can be used to display users' locations on a real-time map (which can also highlight other interesting Web-retrieved information for the group, such as museums or hotels, depending on the specific interests of the group).

It is worth noticing that our current implementation deals with privacy by leaving up to the individual user to decide whether or not to share his/her position (and with which accuracy), make it available only to a restricted group of users or to make it publicly available but only in an anonymous way.

In any case, the results are then visualized at the correct location (i.e., in the form of Google Earth/Google Maps placemarks). Since the answer to a query depends often on the locations of the mobile users, the results of these queries dynamically change as the users change their location in a context-aware way (see Figure 7.3).

7.6 Conclusions and Related Work

In the past few years, several models addressing how to represent pieces of contextual information and information atoms have been investigated, and several infrastructures approaching the concept of "browsing of the world" have been proposed.

In the area of Web technologies, there are a lot of activities in defining proper XML formats (e.g., RSS and Atom) for representing, accessing and organizing information atoms. Unfortunately, these proposals lack by not accounting for the

needs of pervasive computing technologies and context-awareness, e.g., missing standardized ways to represent space, time and activities. Google Earth placemarks clearly emphasize the role of space, though only in absolute physical terms, but still disregard time and activities.

In the area of pervasive computing and context-awareness, most of the focus so far has been on acquiring contextual information and making it available to services for processing in terms of simple key-value pairs (Schilit et al. 1994; Dey et al. 1999). Some recent proposals recognize the need for a more structured model for contextual data, to include, e.g., notions of space (Julien and Roman 2002) and activities (Hong 2002). However, with respect to the W4 model, these proposals appear limited and partial. Two notable exceptions are described in Xu and Cheung (2002) and Bravo et al. (2006), where, with different flavors, the proposal is to structure contextual data around various fields, some of which can be assimilated to our four "Ws". However, these proposals are mainly focused on specific application domains.

From the software viewpoint, several interesting proposals for context-aware browsing of the world services are emerging. Just to make an example, MapWiki (Teranishi et al. 2006) defines a collaborative environment for spreading shared contents on a map in a ubiquitous and location-dependent way, but relies on a centralized approach and does not consider exploiting the presence of embedded pervasive computing devices. On the opposite side, systems such as TinyLime (Curino et al. 2005) and the one described in Mamei et al. (2006) support context-aware access to information produced by surrounding devices (i.e., sensor networks and RFID tags) but totally miss in considering the integration with the Web dimension. The FLAME2008 project (Weissenberg et al. 2004) proposes a general semantic model for providing users with a personalized context-aware access to a variety of browsing of the world services. Still, the model mostly disregards pervasive computing devices and strictly relies on the availability of a centralized information service.

To conclude, we think that our W4 model and the associated infrastructure we have developed represent a significant advance towards the ubiquitous provisioning of effective context-aware browsing of the world services. At the same time, we are also aware they are still preliminary and incomplete. On the one hand, there is the need to better integrate ontologies and semantic analysis in our model, to allow for more semantic forms of pattern matching among W4 tuples and the identification and exploitation of semantic relations between W4 tuples. On the other hand, there is the need to explore more elaborate and flexible strategies for data distribution (including wireless ad hoc communication between users), also with the goal of improving the robustness and autonomicity of our infrastructure.

Chapter 8

Spatiotemporal-Thematic Data Processing for the Semantic Web

Farshad Hakimpour • Boanerges Aleman-Meza • Matthew Perry • Amit Sheth

Abstract. *This chapter presents practical approaches to data processing in the space, time and theme dimensions using existing Semantic Web technologies. It describes how we obtain geographic and event data from Internet sources and also how we integrate them into an RDF store. We briefly introduce a set of functionalities in space, time and semantics. These functionalities are implemented based on our existing technology for main-memory-based RDF data processing developed at the LSDIS Lab. A number of these functionalities are exposed as REST Web services. We present two sample client-side applications that are developed using a combination of our services with Google Maps service.*

8.1 Introduction

Web search is one of the most successful applications in the Internet as exemplified by widely used search engines. The Semantic Web is aimed to improve the capability of such a system beyond a simple keyword search. In the spatiotemporal context, the integration of time and space for searching data sources has been addressing retrieval of the position of entities. With the popularity of spatial data on the Web and the increasing adoption of Semantic Web technologies, the idea of the Geospatial Semantic Web is introduced (Egenhofer 2002). Adding the temporal dimension alongside the spatial and semantic dimensions (Mennis et al. 2000; Perry et al. 2006) increases our analytical capabilities and requires addressing new data integration challenges. This chapter describes our approach to integrating spatial information with event data (i.e., temporal and thematic data) and performing semantic, spatial and temporal analysis on the results. Using spatial and temporal data where available can increase accuracy and efficiency of processes such as disambiguation. The technical contributions of this chapter are in three areas:

- We represent spatial data using the Semantic Web technology (RDF) and enhance this information with spatial relations. We experimented with a geographic data set of the state of Georgia for which we generated RDF metadata representing major geographic features and their topological relations.
- We enrich event data by relating them to associated spatial data. Specifically, we add geographic positions to event descriptions (by geocoding the address of the venues). We also relate address information (street, zip code, state) to the spatial data described above.
- We introduce a set of processes on the spatial, temporal and semantic dimensions of events and show applications built using these processes. Using a set of semantic analytic and event query processing tools, we show how the generated data can be used to build applications.

This chapter is organized as follows. Section 8.2 gives an overview of our data acquisition and preparation including integration issues and disambiguation. In Section 8.3, we present a set of operations for querying space, time and semantics. Section 8.4 presents our experimental systems using the data and operations introduced previously. We discuss the related work in Section 8.5, and Section 8.6 provides conclusions.

We discuss preparation of two types of data in this chapter: first, geographic data from the U.S. Census Bureau[44] and second, entertainment events from several sources on the Web. The resulting data sets are publicly available on the LSDIS Web site.[45]

8.2 Geographic Data

We prepared RDF metadata from four different data sets of counties, urban areas, roads and water bodies. The source of the data sets is publicly available geographic information provided by the U.S. Census Bureau for the states of Georgia and Florida. We enhanced the RDF data set by adding the topological relations between entities. Figure 8.1 illustrates the model in which the data are represented – illustrations of RDF schemas in Figures 8.1, 8.2 and 8.3 are generated by RDF-Gravity.[46]

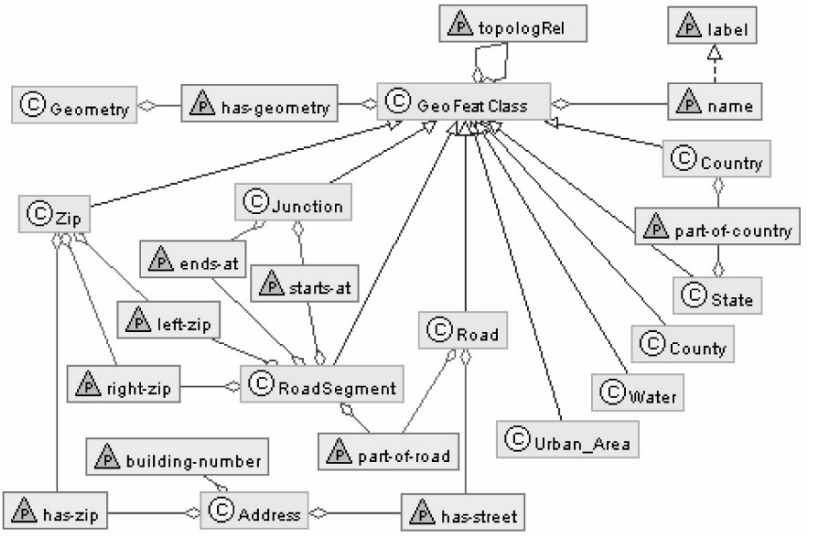


Figure 8.1: The RDF schema for geographic features; the symbols (C) and (P) denote RDF classes and properties, respectively

The main components of this model are as follows:

- The Geographic Feature Class is the superclass of the main geographic entity classes. Entities are transformed to RDF with their corresponding attributes.
- The Geometry class is foreseen in the model to keep the position and shape of geographic features and complies with the OGC Simple Feature Specification (OGC 1999) (Figure 8.2). However, we did not populate our RDF data sets by the

geometry of these objects. In fact, one of our objectives in this work has been that of performing semantic analysis on the spatial objects while relying on existing spatial processing engines (as presented in Section 8.4.1).

- Topological relations are added values obtained using the Oracle Spatial engine (Oracle 2005, Section 1.8) – e.g., relations between zip code and state, county and state, road and county, venues and towns. We use these relations for associating events without keeping the geometries in our RDF data store. As a result, during the process of finding associations we only perform retrieval operations on the topological relations stored in the RDF store.
- Address is a placeholder that can be used in any other data set to relate other objects (e.g., venue in Figure 8.3) to spatial entities, such as zip code, road and state.

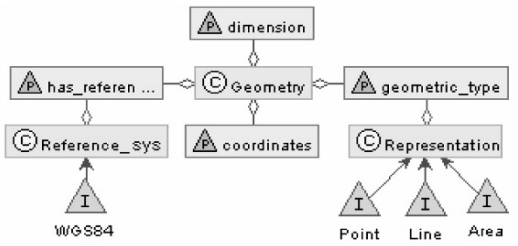


Figure 8.2: OGC geometry model on RDF

8.2.1 Event Data

The event data discussed here are extracted from three different Web sites: eventful.com, atlanta.creativeloafing.com and ticketmaster.com (Table 8.1). For scraping we used the NekoHtml Java library.[47] Different programs crawl and extract events from these Web sites. Data items obtained and modeled for every event include (Figure 8.3)

- event title: a phrase containing a few words briefly describing an event.
- event time: a time point or a time interval. In most of the cases we have only the starting time, because most of the events in our event data sources consistently provide it.
- event location: the venue where the event takes place. The above sites provide users with the information about the venue of every event. The data we extracted related to the venues are as follows:
  - title: presents the name of the venue.
  - address: relates venues to the spatial data (see event model in Figure 8.1). This class is generated according to the extracted data.
  - geometry: keeps the geographic position and the shape of the venue or the event (Figure 8.2). The geometry information is obtained from the Yahoo! Geocoding API.[48] Events are also related to the Geometry class for special cases where an event occurs in a position without a venue, such as an accident.

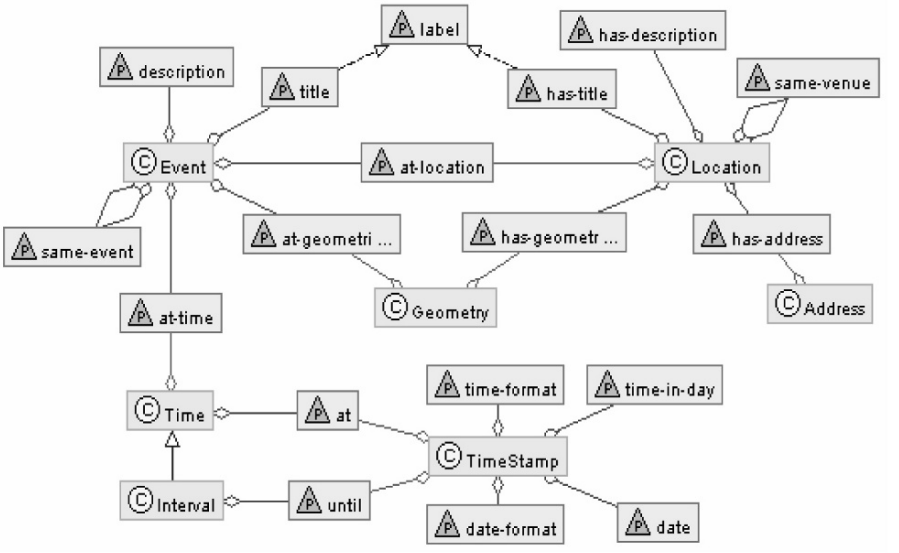


Figure 8.3: RDF schema for events and their time and venue

8.2.2 Data Integration and Disambiguation

Schematic and semantic integration of the data sets obtained from several sources is the next step (Sheth 1999). The schematic integration has not been a major challenge considering flexibilities provided by RDF. Semantic integration, however, presented significant challenges.

Due to the use of several data sources for events and venues, obtaining different event (or venue) resources referencing the same real-world entity is inevitable. This problem is known as the reference reconciliation or entity disambiguation problem (Dong et al. 2005; Tejada et al. 2001). Furthermore, various forms of objects may be incompatibilities or conflicts (Kashyap and Sheth 1996). Such ambiguities are resolved during our integration process.

Existing disambiguation approaches typically rely on either text matching such as (Li et al. 2005) or object attribute matching (Dong et al. 2005; Tejada et al. 2001). Our approach extends traditional methods by incorporating spatial and temporal attributes. We used a combination of two stages of position matching and then title matching for resolving ambiguity of the identity of venues. For events, the disambiguation process is performed in three steps: Time Matching, Venue Matching and finally Title Matching. Figure 8.4 illustrates an example disambiguation process for event E1 by matching it against other events.

Table 8.1 shows the number of events and venues extracted from different sources (in June 2006). The numbers in parentheses show how many venue addresses failed during the geocoding process. The higher number of incomplete or false addresses for eventful.com was expected as the information on this site is entered by Internet users. The last row of Table 8.1 shows the number of unique events and venues after the disambiguation process.

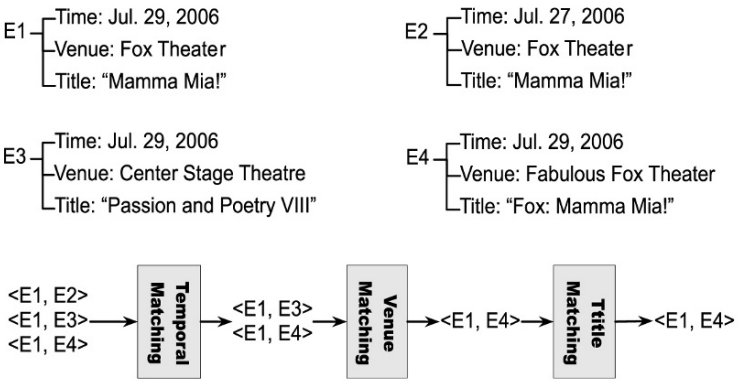


Figure 8.4: Illustration of an example for event disambiguation

During the extraction process, we obtained events that we were not immediately able to classify due to lack of information. However, we are able to improve the event classification by the knowledge acquired from their venues. First, our system assigns usage tags to venues specifying the type of events taking place in a venue. Second, for every unclassified event, the system classifies the event based on the usage tags assigned to its venue. Finally, we created required relations between the address of venues and the geographic features such as roads and zip codes in our geographic data set.

Table 8.1: Number of extracted events and venues

	Events	Venues
creativeloafing.com	10739	807(−18)
eventful.com	1419	717(−284)
ticketmaster.com	311	34(0)
Total	12469	1558(−302)
Total after disambiguation	12156	1267(−302)

As part of the integration process, we relate the event data to our spatial data through addresses. A complete address allows us to disambiguate a street name and resolve it to URIs in our geographic data where possible. In other words, road or street names are not kept as literals but rather by a property from an address to a road instance. The service for resolving a street name of an address to a street URI in our spatial data set is publicly available.[52] An advantage of this process is that it allows us to relate a venue to roads or streets. This facilitates responding to queries such as finding all venues (or events) at a specific street.

8.3 Spatial Temporal and Semantics Analysis

In this section we introduce a set of spatial, temporal and thematic (or semantic) operations we provide on our event data set. These operations are used in our STT (spatial, temporal and thematic) disambiguation process. The main focus of these operations is finding STT proximity in these three dimensions.

We measure proximity in space based on a distance function. Finding the nearest neighbor for a position is a known operator in the spatial domain. We define this functionality by the following operation:

(1) *nearestEvent (type, pos, n)*

where *type* is the type of event of interest, *pos* defines the position for the neighborhood function and *n* defines the number of events in the result list. The result list is sorted by the distance from pos. An example of such proximity query is "finding the closest musical play near my office":

nearestEvent(<musical_play>, <33.946, -83.374>, 1)

We extend the above proximity operation in time as measured through the following two functions:

(2) *nearestEventBefore(type, t, n)*
nearestEventAfter(type, t, n)

where *type* is the event type of interest, *t* specifies the time for the neighborhood measure and *n* defines number of events in the sorted result list. The result of *nearestEventBefore* is in descending order and that of *nearestEventAfter* is in ascending order. An example of such a query is a request to "find 10 speeches right after the working hour on July 22":

nearestEventAfter(<class>, <July 22, 2006, 17:30>, 10)

We use the association ranking developed at LSDIS and introduced in (Aleman-Meza et al. 2005) as a measure for semantic proximity:

(3) *associatedEvent(type, resource, n)*

where *type* is again the event type of interest and *resource* determines an instance in the RDF graph. This function finds an event that is associated to the resource through a path in the RDF graph and returns the ones ranked highest. An example of such request is a query to find a performance involving a particular favorite artist or an event organized by a specific charity organization:

associatedEvent(<comedy_play>, <Reed Martin>, 1)

The proximity operators shown above operate on each of the dimensions. However, one may look for a nearest musical show in both temporal and spatial dimensions. In such cases the nearest neighbors in temporal and spatial dimensions often are not necessarily the same events. For example, an event *e1* is the nearest event in temporal vicinity (one hour) of our requested time and spatial vicinity of 20 miles, while event *e2* is the nearest event in spatial vicinity of our requested location (3 miles) but takes place four hours after our preferred time.

There is a need for a compromise or prioritization to identify a more suitable event in such cases. Using cost coefficients, we define a spatiotemporal nearest-neighborhood position as follows:

(4) *nearestEventBefore(type, t, pos, tCost, dCost)*
nearestEventAfter(type, t, pos, tCost, dCost)

where *type* is the event type of interest, *t* and *pos* declare the point of interest in time and space dimensions, *tCost* is the cost of time difference per hour and *dCost* is the cost of the distance per mile. The above function returns those events that minimize the following cost function:

(5) $cost(e) = (tCost*timeDiff(time(e), t)) + (dCost*dist(position(e), pos))$

It returns a list of events sorted by the cost function. Finally, adding a parameter to the query in (6) for finding an event associated to an entity can satisfy major proximity queries:

(6) *nearestEvent(type, t, pos, res, tCost, dCost, rank)*

An example of such a query is "find a theater play starring a particular actor and taking place close to my office after working hour on 22nd July". However, if the venue is close to the office, one may be willing to wait a day or two, rather than traveling a long way to the neighboring town and joining the event right away:

nearestEvent(<theater_play>,<July 22, 2006, 17:30>, <33.946, -83.374>, <Reed Martin>, 6, 1, 0.2)

By setting *tCost* = 6 and *dCost* = 1, we express the fact that for the cost of traveling 1 km we would wait 6 hours. Finally, by setting rank to 0.2, in fact, we accept most of events that have any association with Reed Martin. Alternatively, an application may wish to bias this cost function to favor time (e.g., it may be preferable to drive 20 miles rather than to go to an event that impinges on the dinner time as far as the event is on the preferred day).

8.4 Sample Applications

This section introduces two applications that work with our data sets. The first application is intended to show how spatial information can contribute to semantic analytic operations. This application is based on a generic semantic analytic tool that finds and ranks semantic associations in an RDF graph. With the addition of spatial knowledge to our data set, this tool can associate events in the spatial dimension. The second application is intended to demonstrate how this integrated data set allows retrieval of event data. This application uses the proximity functions introduced in the previous section to find suitable entertainment events. We enable users to search the event data using the integration of proximity constraints in space, time and semantics. In fact, the new semantic proximity dimension is introduced to the known spatial and temporal proximity dimensions.

8.4.1 Adding Spatial Information to Semantic Analysis

First, we show how spatial relations can enrich semantic associations. In short, a semantic association is a sequence of resources and properties in an RDF graph in a way that from each resource there is one property to the succeeding resource (Anyanwu and Sheth 2003). There can be a very large number of semantic associations between two resources – often much larger than the number of documents that a search engine can find in response to keywords. A simple example showing three different paths associating two events is illustrated in Figure 8.5 (graph illustration is done using the JUNG programming library).[49]

Each of the paths conveys a different semantics based on the semantics of the intermediate edges on the path. One shows that both events are taking place in Atlanta, while the other shows that both have geometries according to WGS84. While the former may be an interesting fact, the latter is an assumption we took for granted while processing the geometric data. Depending on our requirements, we may have different priorities in finding such associations.

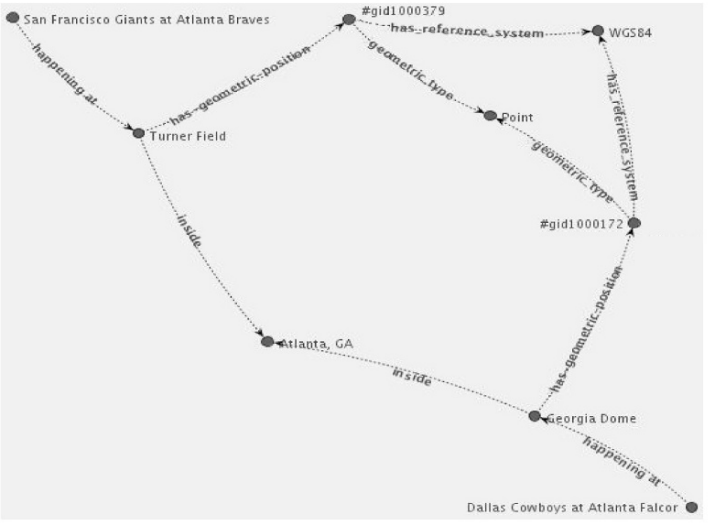


Figure 8.5: Several paths associating two events

This makes the issue of ranking semantic associations very important as well as challenging. Several approaches for finding and ranking these associations are discussed in Aleman-Meza et al. (2005). By means of adding spatial information to entities in the RDF ontologies, spatial objects and their topological relations take part in identifying and ranking the semantic associations.

Figure 8.6 shows how different RDF ontologies can be selected and loaded into the system for finding semantic associations. The ontologies are organized in modules to avoid unnecessary loading of data into memory. For example, if urban areas are of our interest, we do not load the spatial information about counties.

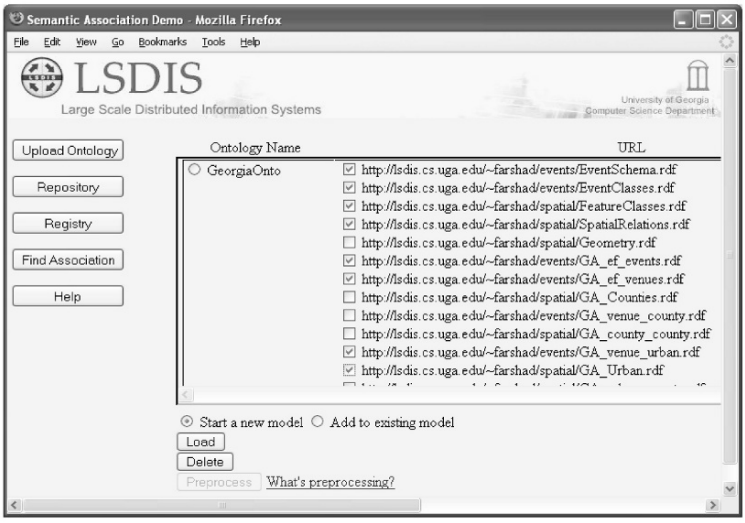


Figure 8.6: Loading RDF metadata sets to find semantic associations

In the next step, we run one of our semantic association-ranking algorithms and also add an ability to visualize these associations. A query to find associations between "Dallas Cowboys" and "Chicago Cubs" results in a number of associations. An association that contains spatial relations is illustrated in Figure 8.7 (left). The association shows that both teams have matches scheduled at venues in Atlanta. As two of the resources in the association are venues and related to geographic positions, we are able to illustrate them on a map. The visualization of the venues in our example path (Georgia Dome, Turner Field) using the Google Maps API is shown in Figure 8.7 (right). The above system is publicly available at the LSDIS Web site.[50]

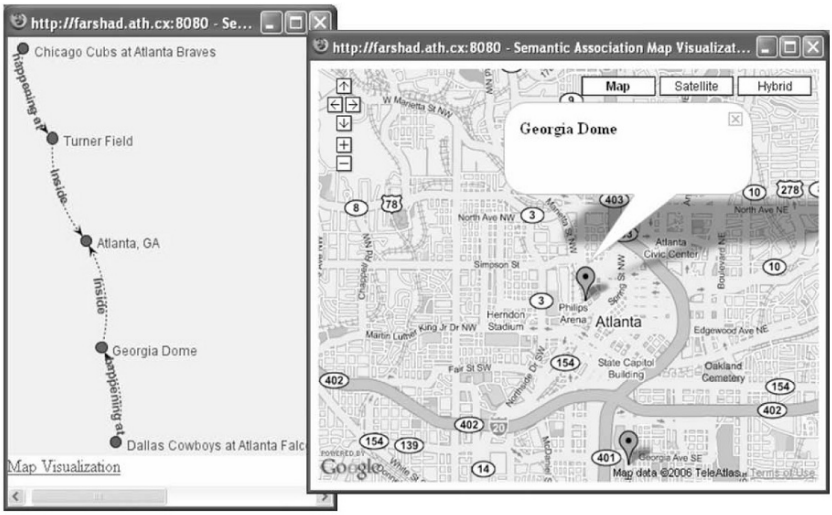


Figure 8.7: A semantic association involving spatial relations on the left; geographic entities in the association are illustrated on the right

8.4.2 Semantics as a Dimension alongside Space and Time

In this section, we show how an application using the functionalities introduced in Section 8.3 is able to find suitable events. As the first step, a set of REST Web services based on the functionalities in Section 8.3 is exposed to the Web. The spatial processing of the operators is implemented using the Oracle Spatial Module (Oracle 2005). We mainly relied on Oracle for its spatial indexing capabilities. For RDF processing we used SemDis API.[51] It is an API for accessing RDF data stores with several different implementations for different purposes. We used the Java implementation for this application. The implemented services are publicly available at the LSDIS Web site.[52] We provide a client-side application that allows a user to specify a set of request parameters. These parameters are used to invoke our REST services as follows:

- time: date and time of day (default: current browser time)
- space: location by specifying an address or by clicking on the map. In the case of entering an address, the client geocodes the address using the Google geocoding service on the client side (unlike the integration process, where we used the Yahoo! Geocoding API) and then sends the position.

- semantics: semantics of events can be constrained in two ways: (1) by specifying an event type, the user can narrow down the type of events; (2) by providing keywords that we relate to the resources in our RDF graph and then associate with the events in our data set.
- costs: cost ratio of time and space. The application provides a slider that helps the user specify the importance of the temporal constraint as related to the spatial constraint. The cost ratio is translated to a verbal sentence describing the preference expressed by the ratio. For example, how much one would be willing to travel to join an event that takes place an hour earlier; or, how long one would wait to travel 1 km less.

Finally, the result of the service invocation is displayed on the map. A snapshot of the client-side user interface is presented in Figure 8.8. The example illustrates a query where a user looks for a theater show nearest to a point she specified on the map and around July 19, 2006. She also indicates an interest in events related to Reed Martin. The system found a show on July 20, 3 km from the indicated point. Furthermore, Reed Martin is both a writer and a player in "The Complete Works of William Shakespeare".

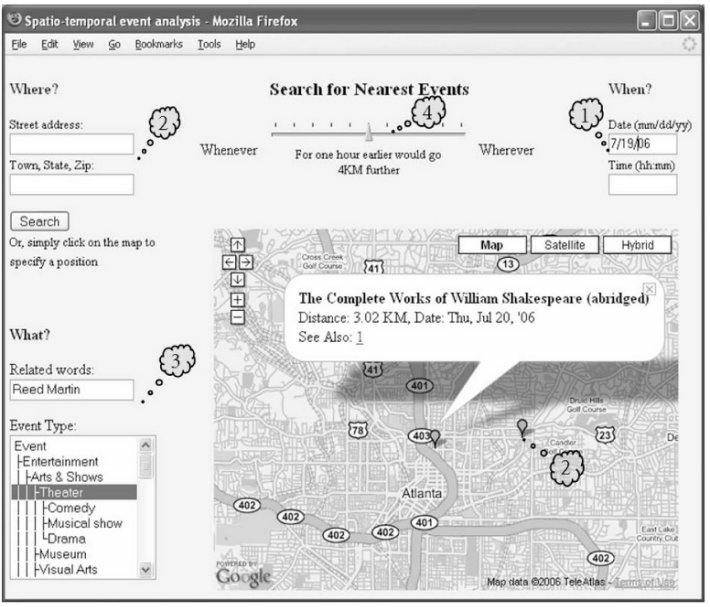


Figure 8.8: A sample application based on our proximity functionality

8.5 Related Work

Our work is related to previous work in different domains, namely, data acquisition, spatial data modeling in RDF, disambiguation and finally event modeling and processing. We used tailored Java code (using the NekoHtml library) for Web scraping, because of the flexibility in generating output RDF data sets and in scheduling of extractors. However, as Semantic Web technologies are gaining popularity, more extraction tools (Hammond et al. 2002) and specifications (Hazaël-Massieux and

Connolly 2006) are becoming available with enhanced capabilities. We believe that in the near future, more RDF metadata will be available as well as better alternative tools for data extraction.

On modeling of spatial information, activities of the RDF community are limited to modeling latitude and longitude of points.[53] We used a more expressive model by adopting the OCG Abstract Specification (OCG 2006c). Another alternative in this area would be adopting GML – a more complex specification. We believe such a level of complexity is not necessary for lightweight spatial processing such as the type of semantic applications discussed here. However, enterprise-centric and scientific semantic applications may benefit from more complex specifications.

Work on disambiguation can be divided into two categories: disambiguation of objects in text as in Li et al. (2005) and disambiguation of objects from different data sets as in Dong et al. (2005) and Tejada et al. (2001). Our work is similar to (Dong et al. 2005) and (Tejada et al. 2001) in the sense that they are also concerned with object disambiguation based on object attributes. However, we take advantage of temporal and spatial attributes of venues and events.

Part of this work is about event modeling and processing. There is a good body of work on spatiotemporal data processing; however, this chapter aims at modeling and processing in semantics, space and time. A similar work in this domain that pays reasonable attention to the STT dimensions is presented by Westermann and Jain (2006). It presents an event-based system for a different domain of application, multimedia information management and a vision of emerging event-based applications.

8.6 Conclusions

The focus of this chapter is presenting our practical approaches for integrating semantics, space and time. Considering that the amount of information related to events is increasing, we explored the integration of spatial, temporal and thematic information from different sources on the Web. We show how information related to the space, time and theme of events can be integrated. The chapter also presents query operators that allow users to integrate constraints on proximity in these dimensions.

We present a description of steps for data preparation and integration. We introduce a subset of proximity operators developed at LSDIS for querying event data. Finally, we discuss two systems implemented using the Web and the Semantic Web infrastructure working with spatial, temporal and thematic data.

Acknowledgements. This work is partially funded by NSF-ITR Award 0325464, entitled "SemDIS: Discovering Complex Relationships in the Semantic Web".

Chapter 9

A Semantic Approach for Geospatial Information Extraction from Unstructured Documents

Christian Sallaberry • Mauro Gaio • Julien Lesbegueries • Pierre Loustau

Abstract. *Local cultural heritage document collections are characterized by their content, which is strongly attached to a territory and its land history (i.e., geographical references). Our contribution aims at making the content retrieval process more efficient whenever a query includes geographic criteria. We propose a core model for a formal representation of geographic information. It takes into account characteristics of different modes of expression, such as written language, captures of drawings, maps, photographs, etc. We have developed a prototype that fully implements geographic information extraction (IE) and geographic information retrieval (IR) processes. All PIV prototype processing resources are designed as Web Services. We propose a geographic IE process based on semantic treatment as a supplement to classical IE approaches. We implement geographic IR by using intersection computing algorithms that seek out any intersection between formal geocoded representations of geographic information in a user query and similar representations in document collection indexes.*

9.1 Introduction

Smart spatial information extraction and retrieval in repositories of electronic documents is the main goal of the work presented in this chapter. The semi-structured and nonstructured data are supported by Electronic Document Management Systems (EDMS) or Library Management Systems (LMS). All these systems aim at providing fast and effective content-based access to a large amount of information. But if we consider that they usually implement statistical approaches to retrieve information, they are insufficient for queries in which the semantics of the search criteria concerns spatial relations (Clementini et al. 1994).

The Virtual Itineraries of the Pyrenees (PIV) project manages a repository of electronic versions of books, newspapers, postal cards and lithographs of the 19th and 20th centuries. It appears that the information is supported by heterogeneous documents but presents many local sources of cultural heritage and denotes various Pyrenean territorial aspects (Cazenave et al. 2004). This kind of repository is still quite unknown. Moreover, it is accessible only in local-area archives of museums and libraries. This is the reason why the regional media library (MIDR) supports this project and intends to diffuse these resources. Thereby the PIV project proposes a semantic approach to analyzing and interpreting geographic information contained in such a corpus or query (Etcheverry et al. 2005; Marquesuzaà et al. 2005). The PIV system proposes to extend basic services of existing LMSs to include new services dedicated to the marking and retrieval of geographic information. It relies on a specific open architecture based on Web services as well as a model describing

geographic information and XML indexes. The originality of our approach is the geographic core model that allows one to formalize any geographic information, regardless of its mode of expression (i.e., text, image). This approach is based on the incremental enrichment of electronic documents. It supports complex processing streams that involve various resource types. The results of each subprocess produce a new XML stream upon which the subsequent subprocesses can rely.

This chapter is dedicated to the marking and indexing of geospatial information. We present related work in Section 9.2 and the PIV geographic core model in Section 9.3. In Section 9.4, we present the PIV prototype; its spatial information extraction and retrieval processes are dealt with in Sections 9.5 and 9.6. Finally, we present an evaluation of this information extraction prototype in Section 0.

9.2 Approaches for Specific Geographic Needs

Information extraction (IE) generally organizes indexes in order to better support information retrieval (IR). Natural Language Processing (NLP) allows specific IE from textual documents, i.e., named entity recognition on diverse types of text (Maynard et al. 2003). Used together, these approaches have the potential to create powerful tools in content-based information systems (Gaizauskas and Wilks 1998).

9.2.1 Information Extraction and Retrieval

IE may be described as the activity of populating a structured information repository from an unstructured information source (Gaizauskas 2002). In a collection of documents, the result of an IE process constitutes what is called an index. It generally consists of a list of terms linked to each document (Tebri 2004). These terms have to describe as precisely as possible the contents of the documents. The automatic IE processes extract either the entire information of a document or only specific parts of it. For example, in the former case, textual processes generally use statistical approaches (all terms of a document are treated) to associate a weight to each term (Zipf 1949). However, in the latter one, they use predefined rules in order to find out specific information (Gaizauskas 2002).

IR deals with models, techniques and procedures to extract information that has already been treated, organized and stored – databases, files, XML files, etc. (Baeza-Yates and Ribiero-Neto 1999). As IE and IR approaches are rather generic, accurate management of spatial information is yet a great challenge.

9.2.2 Natural Language Processing

When dealing with textual documents, a "standard" NLP is based on a set of processing resources (Abolhassani et al. 2003) sequentially applied to all the textual flow: (1) *a tokenizer* is used for splitting the text into tokens; (2) *a splitter* is used to segment the text into logical substructures like sections, paragraphs and sentences; (3) *a Part-of-Speech tagger* produces a POS tag for each token in each sentence; (4) finally, *a semantic tagger* generally consists of grammar sets. Each grammar set contains a series of rules, which act on previously assigned tags in order to produce annotated sets of tokens and/or sentences.

Some systems, like Brill (Brill 1992), Cordial[54] and Tree-Tagger (Schmid 1994), are dedicated resources of such subprocesses. Other ones, like GIPSY (Woodruff and Plaunt 1994), Linguastream (Bilhaut 2003; Widlocher and Bilhaut 2005), SPIRIT (Jones et al. 2004) and GATE (Gaizauskas et al. 1995; Cunningham et al.

1996), support the whole process. Linguastream and GATE are general architectures for text engineering, whereas SPIRIT is dedicated to spatial information. SPIRIT manages a Web document's content and structure, its spatial named entities as well as spatial named entities and spatial relations in queries.

9.2.3 Specific Geographic Needs

According to A. Borillo's (Borillo 1998) hypothesis, natural human language matches a place to a category and associates it with a natural or artificial boundary. In our corpus four main categories can be considered: named boundaries (countries, counties, parishes, etc.), hydrographic features (rivers, estuaries, lakes, etc.), manmade features (cities, towns, villages) and physiographic features (mountains, plains, coastlines). Referring to such places could involve relations with one or more elements; so Vandeloise proposes the principle of the target/landmark pair (Vandeloise 1986). This explains why we understand the first sentence perfectly, whereas the second sentence seems a bit unusual: (1) "the mountain near Pau"; (2) "Pau near the mountain". Our assumption is to extend this hypothesis to any other mode of expression, particularly for maps, lithographs and postal cards.

Early in the study of the computer-based representation of geographic phenomena [in particular, the Geographical Information Systems (GIS) community], geographical information was modeled as a set of geocoded data. That meant that a geographic feature (GF) and all its spatial relations were defined in terms of three primitives: direction, distance and a Boolean set of operations. Freeman (Freeman 1975) defined different spatial relationships (e.g., left of, right of, beside, in front of, near, far, touching, between, inside). However, more recent work has focused on qualitative spatial reasoning (Bennett 1996). Egenhofer and Franzosa (1991) have developed an often-cited model of topological relationships among point sets. Clementini et al. (1994) extended this work by taking into account the dimensions of the intersections. These works have defined the fundamental basis of spatial relations and operations. More recent studies have addressed the formalization of the lack of geographic common-sense knowledge (Torres 2002; Egenhofer 2002); for example, in a digital library context, conceptual links in specific spatial ontology have been proposed (Hill 1999; Hernandez 2005). Finally, interesting research (Hill 1999, 2000) concerns digital gazetteers as a particular form of named places' dictionaries.

But one question has yet to be answered: how could qualitative geographic concepts expressed in textual terms ("*The wood is a couple of crossroads down and to the right from Pau's East exit*") and/or in graphical terms (example of Figure 9.1a) be represented in a formal manner for automatic reasoning and retrieving? Therefore, we have defined a core model to take into account most of these specific needs. This model enables the computation of various operations and transformations for IE processing based on geographic semantics. The result is a high-level representation for in-depth information retrieval.

9.3 A Geographic Core Model

In this model, according to linguistic hypothesis, a GF is recursively defined from one or several other GFs, so spatial relations are part of the GF's definition. The target/landmark principle (Vandeloise 1986) can be defined in a recursive manner. Take, for instance, the GF "north of the Biarritz-Pau line", which can be expressed either graphically (Figure 9.1a) or textually. This feature (i) is first defined by "Biar-

ritz" and "Pau" landmarks that are well-known named places, (ii) the term "line" creates a new well-known geometrical object linking the two landmarks and cutting the space into two subspaces, (iii) an orientation relationship creates a reference for the target to focus on.

In Figure 9.1b it appears that a GF has at least one representation (A) with a natural or artificial boundary. It can be specialized (B) into an absolute (AGF), i.e., named place or a relative feature (RGF). An RGF is defined according to a reference, i.e., a relationship linking at least one other GF (C). The cycle represents the recursive definition.

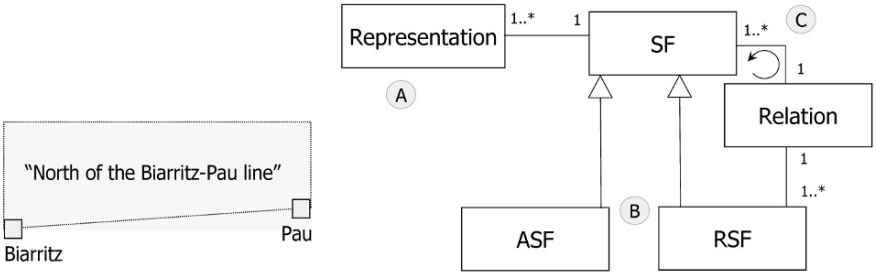


Figure 9.1: (a) GF expressed in a text or a schema; (b) geographic core model simplified schema

Therefore, a GF can be (i) an Absolute Geographic Feature (AGF) if it only consists of a well-known named place, i.e., a toponym with its geocode, (ii) a Relative Geographic Feature (RGF) if it is defined according to a spatial relation (generally topological) linking at least one GF (that can be an AGF or another RGF). For textual IE, this approach has been adapted into a recursive grammar. The following list contains a translated extract of rules originally written in French to markup geographic features:

GF => AGF | RGF.
RGF => relation, GF.
relation => adjacency | orientation | inclusion | distance | geometrical form.
adjacency => "near" | "periphery" | etc.
orientation => "in the south" | "in the north" | etc.
...
AGF => preposition, candidate.

A GF spatial relation could be an adjacency, an inclusion, a distance, a geometric form or an orientation.

- An *adjacency* relation appears when we evoke a GF in relation to its proximity with another GF. This spatial reference is most widespread in written language (example: "near Laruns village" is an RGF, whereas "Laruns village" is an AGF).
- *An orientation* relation appears when we refer to a zone while being directed according to the four cardinal points (Figure 9.5).
- *Distance* appears when we locate a GF by evoking its distance from another GF. Then, we can model the value of this distance and its unit.
- An *inclusion* relation appears when we evoke the inclusion of GF in another area.
- A *geometric form* appears when we need to evoke several GFs to define an unnamed feature with a simple and familiar shape: i.e., "the Biarritz-Pau line".

In the core model, all these spatial references have attributes to characterize them; for instance, *distance* has a numerical or a qualitative parameter, *adjacency* has a qualifier as previously defined in Lesbegueries et al. (2006) and Muller (2002).

9.4 PIV Prototype

The PIV system implements this geographic core model and its geographic features' relations and representations. Figure 9.2 represents the entire processing resources of PIV. Obviously, resulting GFs conform to the core model. Thus, the information extraction subprocesses results are either absolute (e.g., "Laruns village") or relative GFs (e.g., "East of Pau City" or "Laruns village vicinity"). The goal of the PIV system differs from the usual concern, which consists of searching GFs within an information system. First, it must sort through a repository of documents to find GFs that are semantically related to other GFs detected in a free text query. Then, it must extract fragments of relevant documents, classify them and present them to the user. All these tasks are provided by the information retrieval subprocesses (Figure 9.2).

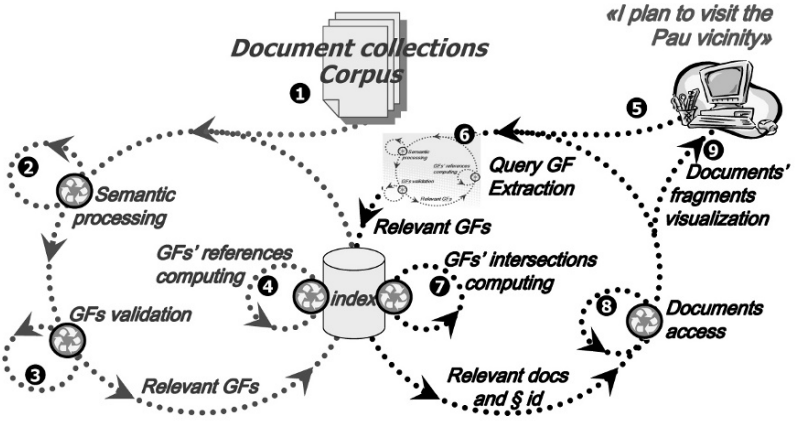


Figure 9.2: PIV system: information extraction (left cycle) and information retrieval (right cycle) processes; subprocesses numbered from 1 to 9 are detailed in next sections

The functions of main information extraction and retrieval processes are implemented as Web services. There are local and remote Web services. For instance, local Web services rely on resources of the French National Geographic Institute. Our linguistic treatment stream is composed of Web services that implement lemmatization, syntactical and semantic analyzers. By using different parameters (grammar rules and lexicons), they might be integrated into other linguistic-specific processes. In the same way, we developed geospatial indexing services supported by a GIS and an XML DBMS. Information Visualization Web services call upon remote services to present results within a cartographic context.

What makes our approach different from other ones like SPIRIT (Jones et al. 2004) and GIPSY (Woodruff and Plaunt 1994) is that it relies on the back-office spatial reasoning used in PIV for the interpretation and indexing of both Absolute and Relative GFs. For instance, the SPIRIT and GIPSY systems only tag AGFs within corpora (Web documents for SPIRIT, specific documents for GIPSY), whereas they tag AGFs and RGFs within textual queries. The PIV system is mainly concerned with domain-specific corpora issued from a source of cultural heritage from a de-

limited and specific region (the southwestern area of France). It implements grammar rules involving spatial relations and geographic literary modes of expression, as well as accurate local specific geographic resources (any fountain, wood or mountain might be validated). Therefore, the PIV system tags AGFs and RGFs in both corpora and queries. RGFs are recursively interpreted. Such an enhanced spatial information semantics interpretation and markup process, within the indexing stage, is made possible because we work on quite stable collections (contrary to Web pages, our digitized documents have to be indexed only one time). So, queries are interpreted dynamically and GF's blow-by-blow indexes allow for more accurate information retrieval. Another specificity concerns the granularity level of the managed information units: textual paragraphs of digitized archives in our case and Web pages in the case of the SPIRIT system.

GFs extracted from various expression modes are managed in the geographic model. However, only the processing of written language has been fully automated for the time being. Sections 9.5 and 9.6 show how the PIV system implements IE and IR complementary approaches (Figure 9.2).

9.5 PIV Geographic Content-based Information Extraction

Hereinafter the subprocesses of Figure 9.2 are described as the linguistic processing sequence supporting the PIV IE process. Its goal is to populate a structured information repository (XML indexes) from a heterogeneous information source.

9.5.1 Data

We collaborate with the Pau County MIDR media library. The corpus (subprocess 1 in Figure 9.2) used for training and testing the PIV system is composed of Pyrenean cultural heritage documents of the XIXth and XXth centuries. This corpus is diverse in terms of modes of expression and genre: (1) textual mode (newspapers and books); (2) iconographic mode (postal cards, lithographs and maps).

9.5.2 Semantic Approach: The Linguistic Processing Sequence and the Validation of Geographic Features

According to work on textual documents (Baccino and Pynte 1994), subprocess 2 in Figure 9.2 adopts an active reading behavior, that is to say, sought-after information is known a priori. This is the reason why, unlike standard natural language processing (Abolhassani et al. 2003), our linguistic processing sequence is locally applied to candidates for named places. To mark these candidates, a "light" lexicon is used in order to have a quite good generic bootstrap process. So AGFs (i.e., villages' names, forests' names, etc.) are detected first and marked. Then RGFs are built by using the previously marked-up AGFs. More precisely, the data processing sequence used to highlight spatial features is implemented as follows:

A tokenizer and a splitter parse the whole of textual flow (Figure 9.3a). This pretreatment corresponds to a new textual flow, where the initial content is added with logical substructure marks or word separator marks added with their lemmas (thanks to an embedded lemmatization phase).

The detection of spatial features called "candidates" is carried out in Figure 9.3b. First, sentences having tokens starting with a capital letter and preceded with a to-

ken containing terms specified in a lexicon, such as "in" and "from" (known as a spatial feature's initiator), are marked. Then, the part-of-speech tagger parses only these marked sentences and retrieves the words' part of speech. For example, in the sentence "Paul passes nearby Laruns", "Paul" and "Laruns" are discerned as proper names ("name" and "pro" values in Figure 9.4a).

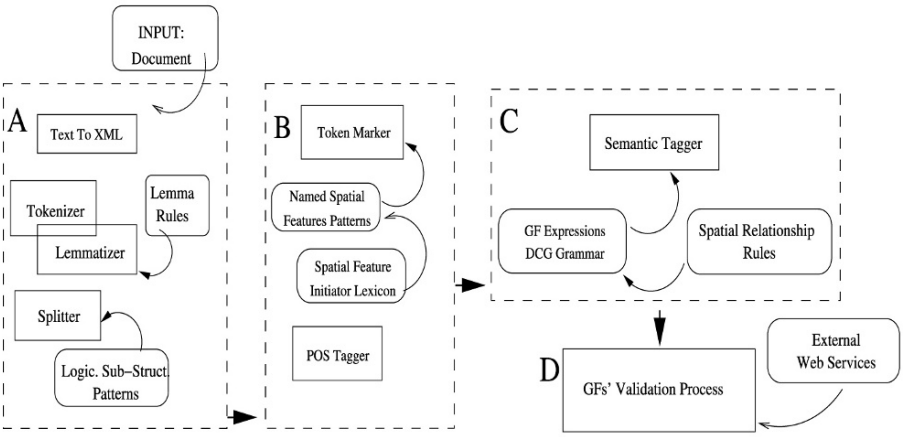


Figure 9.3: Linguistic processing sequence

A Definite Clause Grammar (DCG)-based analysis (Charnois et al. 2003), allowing the interpretation of the extracted syntagms (including inclusion, adjacency, distance to another spatial feature, etc.), is then carried out (Figure 9.3c). The feature "nearby Laruns" is interpreted as an RGF ("rgf" tag in Figure 9.4b), which is itself defined by an adjacency relation and by the AGF "Laruns".

The GF validation stage (subprocess 3 in Figure 9.2) calls external services to confirm every candidate AGF (Figure 9.3d). We use IGN[55] and ViaMichelin[56] resources. For the sentence "Paul passes nearby Laruns", "Laruns" GF is confirmed, whereas "Paul" GF is removed. All the RGFs candidates associated with a nonvalidated AGF are also removed.

During these stages, detection and semantic analysis of geographic expressions are processed on every document, and the results are indexed thanks to an XML semantic markup.

```
<token id="1">
   <text>Paul</text>
   <tag>name</tag>
   <stag>pro</stag>
   <lemma>paul</lemma>
</token>
[...]
<token id="5">
   <text>Laruns</text>
   <tag>name</tag>
   <stag>pro</stag>
   <lemma>laruns</lemma>
</token>
```

A

```
<geographic_feature id="4" id_paragraph="2">
   <rgf>
      <label>nearby Laruns</label>
      <relation>
         <adjacency>
            <type_adj> close to </type_adj>
         </adjacency>
      </relation>
      <agf>
         <label>Laruns</label>
         <type>village</type>
      </agf>
   </rgf>
</geographic_feature>
```

B

Figure 9.4: (a) POS tagger; (b) semantic tagger results

9.5.3 Computing Representations of Geographic Features

GF representations rely on the description of their relations within the Geographic Core Model, as well as on external gazetteers. This final process computes the shapes and georeferences of each GF (subprocess 4 in Figure 9.2).

Representations of Absolute Geographic Features (AGF). If we consider the different levels of granularity and the different levels of precision, the geometrical shape corresponding to the area of an AGF can change. These representations can be built from points (a church, for example), polylines (a road), polygons, multipolygons (a city), etc. We use external gazetteer services to geo-localize AGFs and to compute their geometric shapes. For instance, a district AGF may be represented with at least a geocode. Other representations could be its 2D geometrical boundary or its 3D numerical shape and/or its minimum bounding rectangle (MBR) computed from one of the previous representations (see Figure 9.5, part 1).

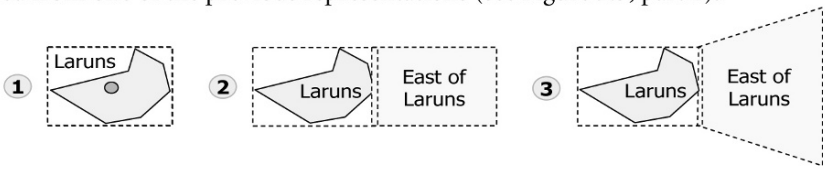


Figure 9.5: Possible representations of Laruns village AGF and east of Laruns RGF

Representations of Relative Geographic Features (RGF). A representation of a RGF like "à l'est de Laruns" (in the east of Laruns village) is computed by a recursive algorithm (Lesbegueries et al. 2006; Loustau 2005). It consists of carrying out recursively geometrical transformations of an AGF and applying some qualitative spatial reasoning mechanisms to it. This algorithm begins by retrieving the minimum bounding rectangle (MBR) of the AGF included in the RGF. Then, it explores recursively the relations that define the RGF and makes geometrical transformations on the original MBR (like translations, homotheties, etc.). Therefore, the "east of Laruns" MBR (Figure 9.5, parts 2 and 3) is computed from the MBR of the "Laruns village" AGF, so a translation on the *x*-axis is carried out. This method may give different representations, and it ensures the proportionality between an approximated RGF and its original AGF. An RGF size is computed according to its AGF size and the semantics of its relations. For geometrical transformations and topological operations, Clementini et al. (1994) have shown that a minimum bounding rectangle (MBR) approach is a quite effective approximation of the shape of the objects.

Index files contain the extracted GFs with their paragraph identifiers (for text documents), their original file identifiers and their geospatial footprints; e.g., for the extracted feature "nearby Laruns" an XML-tree is built and stored in an index file (Figure 9.6). An MBR representation is added to the XML tree (<presentation>).

```
<geographic_feature id="4" id_paragraph="2">
  <rgf>
    <label>nearby Laruns</label>
    <relation> <adjacency>
      <type_adj>close to</type_adj>
    </adjacency> </relation>
    <agf>
      <label>Laruns</label>
      <type>village</type>
    </agf>
  </rgf>
  <presentation> <mbr>
    <xmin>360689.22</xmin>
    <ymin>1752718.63</ymin>
    <xmax>389050.625</xmax>
    <ymax>1789151.375</ymax>
  </mbr> </presentation>
</geographic_feature>
```

Figure 9.6: Example of "nearby Laruns" GF's XML representation

9.6 PIV Geographic Content-based Information Retrieval

9.6.1 Free-Text Query and GF Extraction

A free-text interface supports this stage (subprocesses 5 and 6 in Figure 9.2). Queries are analyzed exactly as the documents of the corpus: the same IE data processing sequence is executed and every GF is extracted. All the validated GFs are geolocalized and a geospatial footprint is attached to each one of these GFs. A query is analyzed online while corpus documents are analyzed offline.

9.6.2 Query: Computing GF Intersections

Our search technique is based on spatial mapping between the GFs of the query and those of the documents (subprocess 7 in Figure 9.2). This mapping is done thanks to the geospatial footprints created dynamically for the query and those stored in the index files of the corpus.

Figure 9.7a illustrates a query and some indexed areas (precise geospatial footprints for AGFs and approximated MBRs for RGFs) that represent Pyrenean villages named in the corpus. Figure 9.7a points out that "Laruns" village is more relevant for the query than "Louvie-Soubiron" village. In the same way, one can deduce that "Gentiane Peak" is not relevant to the same query since its footprint does not overlap with the query one.

The selection process consists of processing index files and computing intersections (Lesbegueries et al. 2006) with a GIS. Then we select corresponding relevant document fragments. We are able to calculate the relevance of a document fragment by computing an evaluation of the surface, which results from the intersection between the GF of the document fragment and that of the query. For any query, the relevance of each recovered document may be different (Figure 9.7b): Df precision = I surface/Df surface; Df significance = I surface/Q surface; Df distance = d/D.

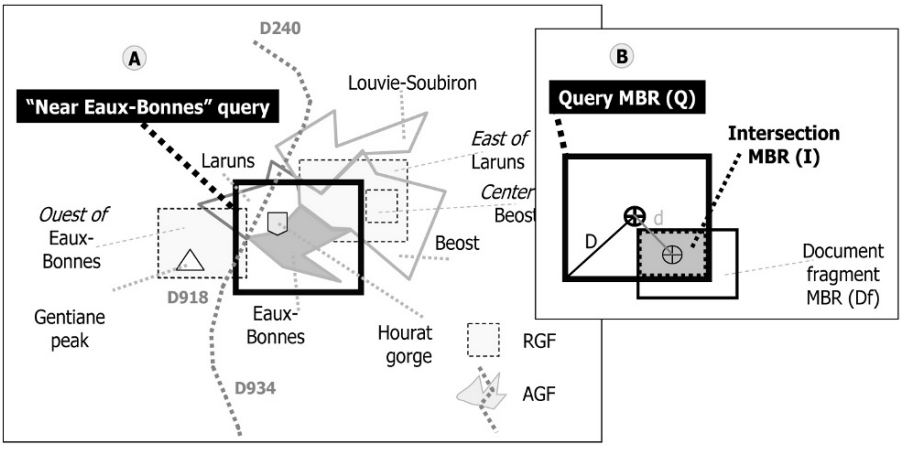


Figure 9.7: (a) An example of query "I want documents dealing with places near Eaux-Bonnes" and its corresponding MBR (the biggest one); the other polygons represent GFs (extracted from documents of our corpus) that may match the query; (b) computing result relevance

Therefore, we compute the Df score = (Df precision + Df significance)/(2 + Df distance). The closer the centroids of I and Q are to each other, the higher the relevance score of the Df. A GIS[57] supports these searching and computing operations on the corpus indexes.

9.6.3 Access and Visualization of Documents

A selected GF's content is reached through XML indexes (XML DBMS),[58] where paragraph identifiers (for text documents) and original file identifiers are stored (subprocesses 8 and 9 in Figure 9.2). IR results are displayed in a Google-like presentation of the relevant paragraphs – GFs are highlighted. We are proposing interactions and interfaces dedicated to such result sets such as visualization, navigation and processing. We also base visualization scenarios on cartographic resources and document aggregation metaphors (Sallaberry et al. 2006). The next section presents experimental results (Figure 9.2, IE process). IR evaluation is in the pipeline.

9.7 Evaluation of the Information Extraction Process

To test and evaluate the geographic core model using different expression modes, some Pyrenean lithographs have been marked manually (stages 1, 2 and 3). Note that the end of this iconographic data processing sequence (stage 4: GFs' geo-localization) is automatic. Consequently, the core model represents GFs for any mode of expression (text, image, etc.), since the common denominator is their geometric form and geo-localization.

Sample data. We carried out scanning and OCR processing of 10 books of our corpora. Then we ran the PIV prototype automatic information extraction processes. The processing of a book of 200 pages (stages 2, 3 and 4) takes five minutes. The PIV prototype found 9,835 candidate GFs in these 10 books. At the same time, we annotated geographic features like in CLEF[59] campaigns, where participants marked GFs of the samples of all these books manually. Finally, we compared these handwritten annotations to PIV results in order to compute the recall and precision rates of PIV IE processes.

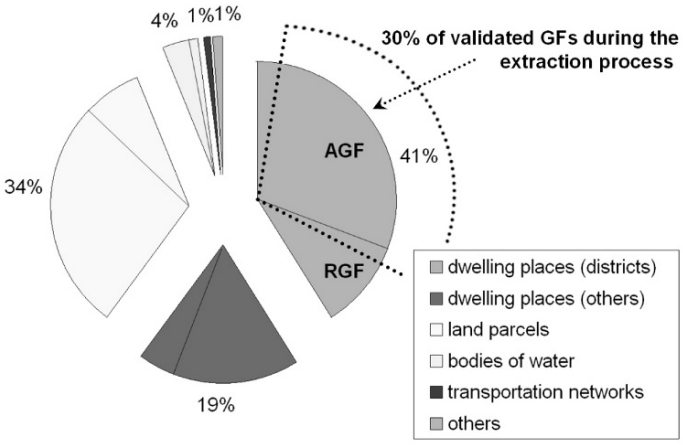


Figure 9.8: GF sample breakdown

Analysis of the data. Linguistic modules of the PIV IE process detect various GFs such as named places of different types of dwelling places (districts or streets, boroughs, counties), land parcels (mountains, valleys), bodies of water (rivers, falls), transportation networks (roads, stations), etc. Figure 9.8 breake down the GFs of the studied samples into these categories.

Evaluation. This study focuses the evaluation on geographic features restricted to districts (towns and villages) – average recall and precision rates equal 49 percent and 73 percent, respectively. These rates are interesting for absolute GFs and quite bad for relative GFs. As further explained in stage 3, they are obtained with few DCG rules (about 10) and validated by named district resources only. In fact, the sampling process allowed the computation of numerous AGFs but fewer complex RGFs. Nevertheless, our use of Definite Clause Grammar detection and the marking of RGFs must be extended in order to overcome this gap. Stages 1, 2, 3 and 4 (Figure 9.2, information extraction process) are analyzed and illustrated in Figure 9.9.

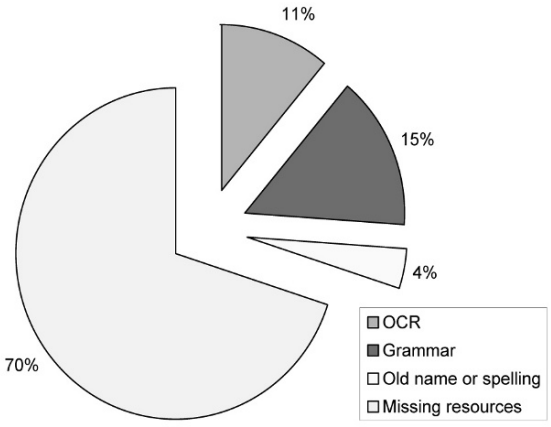


Figure 9.9: Analysis of GF extraction problems

- **Stages 1 and 2: Semantic processing.** Noise is weak; 80 percent of the GFs are detected and marked. Twenty percent of the GFs are not detected because the syntagms responsible for their introduction are not spelled correctly, either because they are not introduced by geographic introducers or because they have no initial capital letter. RGFs like "east of Pau" are captured, whereas more complex ones like "the wood is a couple crossroads down and to the right from Pau's East exit" (cf. Section 2.3) are not detected. More accurate DCG rules (currently being developed) will largely improve this stage. These are, for example, rules dedicated to inclusion relations, while other rules concern skipping the stop words that separate geographic introducers from their corresponding named entity. Stage 2 may also be improved thanks to improved scanning and better OCR tools.
- **Stage 3: Validation.** We validate about 30 percent of the detected GFs. Seventy percent of the GFs cannot be validated since we don't have adequate geographic resources. In fact, we only have resources about district-type dwelling places. Another set of GFs is also lost because they are badly "OCR-ized" or spelled using old French. Stage 3 may be improved thanks to the extension of the resources to other GFs categories. A lexicon of old French spellings of GFs might also improve these results.

- **Stage 4: Computing references.** The AGFs geo-localization process and RGFs approximation (thanks to the MBR approach) are supported by the same Stage 3 resources. So, no GF is lost here.

9.8 Conclusions

We focused this work on restricted and stable corpora such as local cultural heritage collections of documents. This specific context makes it possible to implement sensible scans, which take into account the document content. Our contribution is complementary to the traditional search methods used in libraries. Our objective is to treat the geographic semantics of such collections of documents and users' queries more accurately. The PIV prototype implements an innovative geographic semantics approach on information extraction and retrieval. Its experimentation with heterogeneous (texts and images) document collections shows that this approach improves the relevance of geographic query results (Lesbegueries et al. 2006). The most specific aspects we propose are

- the geographic core model description: it supports a formal description of every geographic feature (GF) detected in collections of texts or images;
- the smart extraction process of textual GFs: this automated IE process first looks for any named Absolute GF and then runs a semantic analysis to find out if a larger Relative GF containing this AGF exists;
- any GF's geo-reference can be found in gazetteers or approximated by using spatial relation semantics interpretation and the MBR approach: the IR process is based on intersection computing between geospatial footprints extracted from a user's query and those of indexes of document collections; the PIV system relies on a Web Services architecture and supports full XML format (schemas, GFs representations, indexes, documents extracts, etc.).

The IE process has been evaluated; the IR process evaluation is ongoing. As pointed out by the first results, main drawbacks concern Definite Clause Grammar rules, misspelled scanned or old place names, and the lack of larger geographic resources. Our work now focuses on extending grammar rules in order to improve the Relative GF capturing process. We are also integrating a new set of spatial resources describing Pyrenean roads, rivers, woods, valleys, mountains, etc. These resources coupled with text management Web services (stem, sound, fuzzy comparison) should consequently improve the PIV system's information extraction and retrieval results.

Future works will concern sets of Geographic Feature summarization. We propose different geographic patterns (itinerary, point of view, area comparison) and qualitative and quantitative characteristics (Lesbegueries et al. 2006b). This approach computes summarized geospatial indexes for different granularity levels of text units (sentences, paragraphs, groups of paragraphs, pages, etc.).

As PIV users are expected to be tourists (Etcheverry et al. 2006) and scholars (Nodenot et al. 2006), we are also adapting the first version of the user interface. Moreover, we are developing a cartographic visualization mode, which will better take into account the territorial specificity of our documents (Sallaberry et al. 2006).

Acknowledgements. Our project is led in partnership with the Greater Pau City Council and the MIDR media library. We want to thank them for providing us with their digital corpus and their support.

Chapter 10

Enhancing RSS Feeds with Extracted Geospatial Information for Further Processing and Visualization

Marc Wick • Torsten Becker

Abstract. *Internet users are flooded with information and are thankful for help in categorizing and visualizing textual content. Geographical categorization is one of the most important criterion for filtering, grouping and prioritizing information as users are naturally more interested in local information. We describe a way to extract geographical information from textual content using natural language processing, and we display the information within a geographical context on maps and satellite images. Using the widely supported RSS format as the input format, this approach allows us to process content from nearly all online news sites and blogs.*

10.1 Introduction

Extraction of geographical information from natural language text is the automated process of analyzing text such as a newspaper article or a blog entry and finding the geographical context of the text. In this case study, we describe a system that reads text in the widely used RSS format, extracts geographical information and finally displays the text items on satellite images and maps. The system is freely available on the Internet and can be used easily to extract and display geographical information from most Internet news platforms and blogs.

The fast-growing Internet is abundant with unstructured information, and it is more and more difficult for human users to find the information important and relevant to them. An important factor in ranking and ordering the plethora of information is locality. News items, blog entries and Web pages often have a geographical point of reference, and the closer the reader is located to this point of reference the more important and relevant is the information to this particular reader.

Automatically extracted geographic information can be used to filter, prioritize or categorize news by software agents and thus help the human user save time and find relevant items of information.

10.1.1 RSS Feeds

Our system is using RSS as the input text format and may also produce RSS as the output of the processing. RSS is a well-defined format for textual data and is meant to be processed by machines in contrast to other text formats like Web sites or word processing documents intended for a human reader. Most Internet news platforms and blogging software offer at least part of their content as RSS feeds, which contain a title and a description or a short abstract of the most recent entries. All this information is therefore easily accessible over the Internet to our natural language geocoder without requiring any action on the publisher's part.

As the system not only consumes but also produces RSS, it can be transparently put between the original producer and the consumer and enhance the original text entry with semantic geographical data such as latitude and longitude. Typical RSS consumers are feed aggregators that are used to create "personal newspapers" and help manage the overwhelming news stream.

10.1.2 Gazetteer

The geographical locations are determined by the use of the comprehensive and freely available geonames.org gazetteer. With over 8 million toponyms, the geonames gazetteer accumulates geographical information from various sources.

In the information extraction process, several different steps can be distinguished. In a first step, each word is tagged with its "part of speech" and assigned a weight before being handed over to the search engine. A scorer then calculates score values for geographical entities returned by the primary search. The toponyms are ordered by their score values, and the best place name is returned if its score is above a minimal threshold.

The input text's language is an important factor in natural language processing, and the processor currently supports five languages. Geographical latitude and longitude are not the only properties of the returned toponym. Others are population, elevation, feature class as well as a feature code, which describes the type of a toponym like mountain, lake or building. The result of the information extraction is made available in several output formats and includes the input source's original information.

10.1.3 Visualization

Visualization of textual information in a geographical context displayed on maps or satellite images helps users enormously in reading and understanding text. Recent developments of free geospatial platforms like NASA Worldwind, Google Earth, Google Maps, MSN Virtual Earth and many many others together with the increase of broadband access have made geo-browsing very popular. Our information extractor makes it possible to read any RSS feed on geospatial platforms.

The processor also supports many RSS dialects as output format and adds latitude and longitude in the newly created GeoRSS[60] namespace, which is already supported by a couple of RSS readers and aggregators. At the moment the most sophisticated geobrowser for RSS feeds probably is the ExploreOurPla.net platform. ExploreOurPla.net is a Web browser-based geographical desktop application with hundreds of data and image layers. Near real-time satellite images allow the news reader to directly see the effect of news events like hurricanes, volcanoes, forest fires and many others.

10.2 Extraction of Information

10.2.1 Geonames.org Gazetteer

A comprehensive gazetteer is fundamental in order to extract geographical information from natural language text. The geonames.org geographical database is available for download free of charge under a creative commons attribution license. It contains over 8 million geographical names and consists of 6.2 million unique fea-

tures, whereof 2.2 million are populated places and 1.8 million are alternate names. All features are categorized into 9 feature classes and further subcategorized into 645 feature codes.

Geonames is aggregating data from various free sources such as the Geographic Names Information System from the U.S. Board on Geographic Names (USGS 2005), the National Geospatial-Intelligence Agency (NGA 2006), geobase.ca, a Canadian governmental initiative, Wikipedia and many more. Other sources like the gtopo30 digital elevation model from the U.S. Geological Survey and population data are used for error correction and data enhancement. The geonames data set is displayed as overlay over a satellite image view to give its users an impression of data coverage and data accuracy. A Wiki interface allows users to manually or programmatically correct errors and insert missing records.

Updates of original data sets are loaded periodically into the geonames database. The NGA, for example, releases a revised data set approximately once per month. During the load of the revised data set into the geonames database, it is made sure for every record that no geonames user has modified the same record. If a conflict is detected, the change history for both update branches is compared and the two branches are merged whenever possible. If it is not possible to reconcile the two conflicting branches, then the modification by the more trustworthy source takes precedence. Not only is the geonames database available as free download, but geonames is also providing a huge number of Web services to directly access the database online and integrate it into other applications. Among these Web services are search or geocoding services where latitude and longitude are returned for a place name or postal code. Reverse geocoding services take latitude and longitude as input and return a feature like a city name, a street name or nearby Wikipedia articles.

10.2.2 RSS to GeoRSS

As data format of the input text for the information extraction, we use the widely used RSS/Atom format. RSS is an acronym for "Really Simple Syndication" or "Rich Site Summary" and is a Web content syndication format. It was first designed and used by Netscape in 1999 and basically is a list of items in XML format with a title, a description and a link to the original Web page of this news item. Nearly all Internet news and blogging platforms provide at least part of their content in so-called RSS feeds, which usually give summaries of the latest news items with back links for detailed information. The RSS format is not directly targeted at a human audience, but it is meant to be consumed and processed by software agents before being handed over to the human reader.

News aggregators reduce the time and effort needed to regularly check Web sites for updates, creating a unique information space or "personal newspaper". The reader no longer has to periodically check dozens of blogs and other Web sites for updates; the aggregator software does this for the user and bundles everything the user is interested in in one place. We believe the next generation of RSS reader software will have a certain kind of intelligence and help users group, filter and prioritize various news items based on their personal interests. These intelligent RSS readers will help users focus on news items relevant to their particular interests. An important criterion for intelligent aggregators will be the geographical location of the user and the news item. The nearer the location of a news item is to the location of the user, the more relevant it is to him or her.

Our natural language processor takes an RSS feed as input and returns the processing output again in RSS format. The original RSS dialect is extended with geo-

graphical information in GeoRSS format. The output returned by the processor can in turn be used as input by RSS readers and aggregators to be displayed or for further processing steps. It is also thinkable to add several processors in a row, each taking RSS as input and producing RSS as output.

10.2.3 Extraction of Geographic Information

For the information extraction, the title and description of each entry in the RSS feed are combined into a single text and tokenized at word boundaries by a simple tokenizer algorithm. It has proved advantageous to set the description in front of the title to simplify the identification of persons' names in the disambiguation of geo/non-geo entities. The title is often using short versions of names to refer to people, whereas the description is using the full name more often. Therefore, it is natural to analyze the title only after having analyzed the description and use the names of people already identified as a reference in analyzing the title.

The tokens created by the tokenizer are then handed over to the part–of-speech tagger, which assigns to each token its most likely part of speech (Manning and Schütze 1999; Jurafski and Martin 2000). For our purpose of extracting geographical information in real time, a full part–of-speech tagger is not required and would need far too much processing time. We therefore use a partial part-of-speech tagger focusing on identifying names of persons or names of nongeographical entities, geographical prepositions, adjectives with geographical meaning, definite articles and indefinite articles. The part-of-speech tagger is using word frequencies from different sources. The U.S. Census (U.S. Census 1990) first- and last-names' frequency tables are used to identify names of persons in English. Wikipedia is a large corpus of natural language text freely available in countless languages and was used to derive frequencies of common words.

After the part-of-speech tagger, a query generator discards irrelevant tokens, assigns weights to relevant tokens and generates a query to be run against an inverted index of the geonames geographical database using the lucene search engine library (Cutting 2006). This inverted index of 8 million toponyms has a size of over 3 GB and holds information like "how many times does a term occur" (term frequency) or "in how many toponyms does a term occur" (inverted document frequency) (Spärck Jones 1972). The information of the inverted index together with the weights of terms in the search query will return a preliminary list of places ordered by a first score.

A scorer process then calculates a second score value for each toponym returned by the search over the inverted index. This final score is based on the relation of different place names returned by the search, since we may have place names in the text used to uniquely identify another place name. In the text fragment "... in Montfermeil, north of Paris ...", the place "Paris" is more relevant then the place "Montfermeil", but the former is used to identify the latter, and consequently the place "Montfermeil" in close distance to Paris will receive a higher score than "Paris". The semantic information of the term "north" is not yet evaluated by the processor, and this will be an interesting improvement for future versions of it.

In "... to Florence, Colorado, ...", the name "Colorado" is used to identify which out of circa 1,700 Florences in the geonames database the text is referring to, and it will be used to increase the score of the toponym "Florence" in the U.S. state of Colorado. The well-known Renaissance city "Florence" in Italy normally has a significantly higher score than the rather small town in Colorado. With the apposition "Colorado", however, the small town in Colorado will score higher.

For the geo/non-geo place name disambiguation, the scorer therefore has to understand how different place names in the text are related to each other. If a place name is used to describe or identify another place name, its score value has to be attributed to the place it is describing. An additional help for the scoring and place name disambiguation is the information already extracted from previous items in the feed. If the previous items fell within a certain geographic area, then a place within this area will score higher then a place distant from the previous items.

After the scorer has calculated a score for each of the most important place names returned by the search, the places are sorted and the best scoring place name will be returned, if its score is above a minimal confidence threshold. The final output of the processing is available in several formats. Many RSS and Atom dialects are supported, combined with the three encodings for geographic information. Further supported output formats are Javascript Object Notation (JSON) for fast mashups and easy integration into Web applications and KML for Google Earth. The JSON encoding includes latitude and longitude as well as additional semantic information like country code, feature class, feature code and many more. In RSS/Atom only latitude and longitude are encoded since none of the three GeoRSS encodings allows encoding semantic information.

Important parameters for the part-of-speech tagger, the assignment of weights for the search and the scoring are the language and the context of the text. The term "Java", for example, in an IT or job context is most likely referring to the programming language, whereas in another context "Java" has a geographic meaning and refers to the Indonesian island. If these parameters are not set explicitly, then the system will use two rudimentary detection algorithms.

The processor currently supports five languages, English, German, French, Italian and Spanish.

In order to get an idea of the accuracy rates achieved by the information extraction, we have used random samples from a corpus of 30,000 news items from Reuters, Yahoo! News and BBC. Preliminary rounds of these tests have shown accuracy rates of about 90 percent.

10.2.4 Related Work

A lot of work is being done in the fields of natural language processing and information extraction, and we can cover only a few projects here. A gazetteer is at the core of every system, and most projects work with homegrown and not freely available or even commercial gazetteers. The geonames project is an effort to provide a free gazetteer, and we hope other teams will join this effort.

Li et al. (2003) have been using InfoXtract to extract geographic references with pattern matching driven by local context, a minimum spanning tree algorithm for place name disambiguation and default senses heuristics. A weighted graph is constructed for place name disambiguation, where each node represents a location sense and each edge represents similarity weight between place name options. Prim's algorithm for calculation the minimum-weight spanning tree has turned out superior to Kruskal's algorithm.

In the G Portal (Zong et al. 2005) digital library project the assignment of place names to Web pages is divided into three subproblems: place name extraction, place name disambiguation and place name assignment. Place name extraction has been accomplished with the use of the entity extraction library "GATE". Three steps build the place name disambiguation process. A first set of place names is determined by self-feature pattern matching. Place names for which a parent place superior in the

administrative hierarchy is found in the local context are disambiguated in the second step. In the third and last step, all remaining ambiguous places are disambiguated with calculating the spatial distance between all place options and then choosing the most adjacent place name. For the place name assignment, the parent places receive scores from their children and the highest scoring place is returned. An often-mentioned place name may get a higher score than its parent.

Web-a-Where (Amitay et al. 2004) is, like G Portal, a system for associating geography with Web pages. It extracts place names from Web pages and assigns each page a geographic focus. Place name disambiguation is done in three steps. First, tokens in the vicinity are used to identify a place name. Then a default sense is used for the remaining ambiguous places. In the third step, places with a default sense below a minimal threshold are disambiguated using the context of the administrative hierarchy of the place name options. In Web-a-Where, the default sense is one of three steps, whereas in the G Portal project, described above, spatial distance is used as the third and last step. The pattern matching in the first step used by G Portal is probably similar or even equivalent to the "vicinity" approach in the Web-a-Where project. Place name assignment is also based on the administrative hierarchy of the disambiguated place names. Up to four foci for a page are found. The paper reports correct geotagging of 80 percent and correct place name assignment of 91 percent.

The SPIRIT Spatial Search Engine (Jones et al. 2004) is a Web search technology for geographical information and consists of the following components: user interface, geographical and domain-specific ontologies, Web document collection and the core search engine. The geo-ontology supports functionality for disambiguation, query expansion, relevance ranking and metadata extraction. Geographical place names are translated into geometric footprints and qualitative spatial relationships. Spatial indexing of documents has been integrated with text indexing through the use of spatio-textual keys in which terms are concatenated with spatial cells to which they relate.

10.3 ExploreOurPla.net: A Geo Blog about Living Near the Coast and Rising Sea Levels

10.3.1 A Challenge for Everybody and the Planet

The world is facing a challenge with the potential of changing every single aspect of human life. It is time to inspect the way we work, travel, dwell and consume. Every nation will suffer from uncontrolled global warming. Rising sea levels, extreme weather conditions or the power of tropical storms could turn places millions of people are living into inhabitable regions, with global migration as a result.

Scientists say there are 10 years to fulfill a turnaround. After this period consequences are unpredictable or unmanageable. The problem is no longer considered as a political one; it concerns any of planet Earth's 6.5 billion inhabitants.

Global problems of this quality need global communication to spread the message. The Internet is available and popular in all developed countries. Emails, Web sites, news groups and newsletters are electronic forms of communication. Everybody with an Internet connection can publish his or her thoughts in a daily blog with a worldwide availability.

10.3.2 Of Linking Content with Events, Near Real-Time Maps and Real Places

ExploreOurPla.net was founded in early 2006 to take part in the run for better climate conditions and less pollution. As blog with the subline: *About living near sea and rising sea levels*, it collects daily news regarding recently published research, polls about environmental topics, campaigns with the goal to reduce emissions and extreme weather phenomena.

To give users a picture about Earth, an interactive map interface based on Google Maps was integrated. The so-called Explorer covers the planet, but due to the chosen Mercator projection, the poles are not exactly displayed. The *Explorer* not only manages the three map layers provided by Google, but adds nearly 20,000 layers from publicly available OGC WMS layers. The range of layers covers topics like population, bathymetry, shaded reliefs, volcanoes, mean temperature and vegetation. All these layers are usable as base map and overlay with transparent or translucent areas. This gives users the opportunity to combine maps and to extract new information.

On top of the maps users may choose from the list of available near-real-time services to display up-to-date information. The data come from public services and are mostly provided by geonames.org. The three most used extensions are METAR Weather with current temperature, humidity and pressure from 5,000+ airports, NRT Earthquakes from USGS with data about magnitude, depth and time and tropical storms with observations about storms' intensity, speed, direction and place. At anytime the place names extensions helps users to keep overview when the zoom level is too close to identify known geographic features.

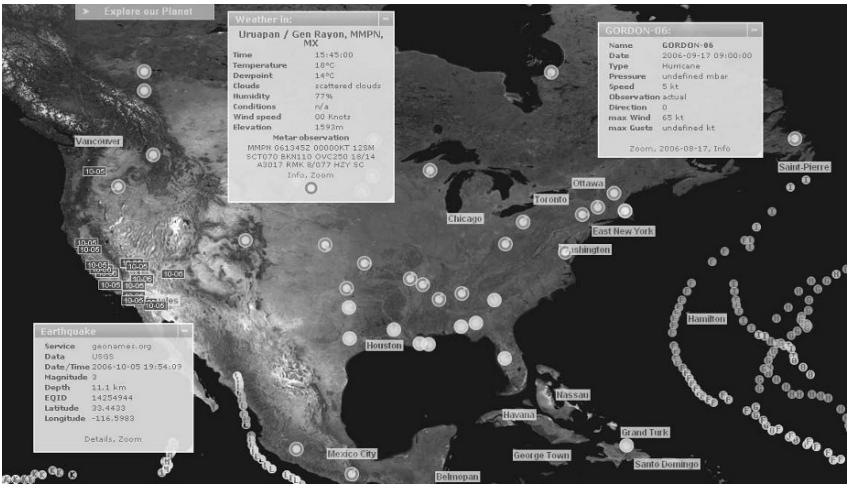


Figure 10.1: Map interface with near–real-time extensions

ExploreOurPla.net implements geoLinks, unique URLs with all parameters needed to drive the map interface. The current base map, all active overlays, extensions, tools and the map date are coded into this permalink and makes it suitable for electronic communication via Internet or as a bookmark. Each blog posting accepts a set of corresponding geoLinks. The blog category *Earth Views* collects posts with time-enabled geoLinks and links to visible events like hurricanes or wildfires with

Daily Terra or Daily Aqua as base map. The MODIS Rapid Response system NASA service has provided middle-resolution (250-m/pixel) WMS layers on a daily basis since December 25, 2004.

Figure 10.2: Map interface with active news and geoFeedExplorer

To link the latest news with places on Earth, the geoFeedExplorer was developed. Like the other extensions it is triggered via hotkeys but has some configuration options for registered users of ExploreOurPla.net. The design goal was to have a map where all news items are displayed as icons, and with a single click the corresponding entry is displayed inside a movable window and readable to the user.

10.3.3 AJAX – Updating without Refreshing the Browser

Four parties are involved in the process of exchanging and preparing the data needed to display a geo-enabled RSS feed to the user. First, the user starts selecting the wanted feed. He may enter the feed's address as parameter:

```
http://ExploreOurPla.net/georss/http://ExploreOurPla.net/feeds/
```

which loads the GeoRSS feed from ExploreOurPla.net without involving the geonames.org Natural Language Geocoder. The following URL loads the news feed from Reuters and turns it into a GeoRSS feed via geonames.org:

```
http://exploreourpla.net/georss/http://today.reuters.com/rss/worldNews
```

To transform a usual RSS feed into a GeoRSS feed, the user may also choose one from the geoFeedExplorer. In any case the Web interface of ExploreOurPla.net (EOP) accepts the URL and sends it to the php proxy on the same domain. The script then calls the GeoRSS service at geonames.org and awaits the response. A timeout may occur here, but the user will be informed about it and has the choice to refresh or restart the request.

The script accepts any valid answer in JSON format and adds some meta information like service name, time needed and size of data. Before sending the JSON

literal object back to the interface, it is wrapped by a CDATA element of the containing XML document. This ensures the transported data are valid and will be accepted by the browser's XMLHTTPRequest object.

Using XML as the unified data format between client and server has the advantage that all kinds of data can be easily interpreted by the same code library, whether JSON, XML or pure text. The JavaScript communication object simply uses the method responseXML.getElementsByTagName() of the DOM and selects metadata and pure data. Only the content data element is then delivered to the geoFeedExplorer.

10.3.4 Combining Map and Content to a New View

The geoFeedExplorer is a container for a three-level hierarchy. The root stands for the geoFeedExplorer itself and contains feeds symbolized with their address or, if loaded with title, colored RSS icon and Web site icon. Most feeds have a description, and each feed has no, one or more entries with different data. It is up to the feed publisher to use and fill the available elements. Depending on the promotion strategy, there are entries with title only or else full-featured news items with source, author, date, title and description.

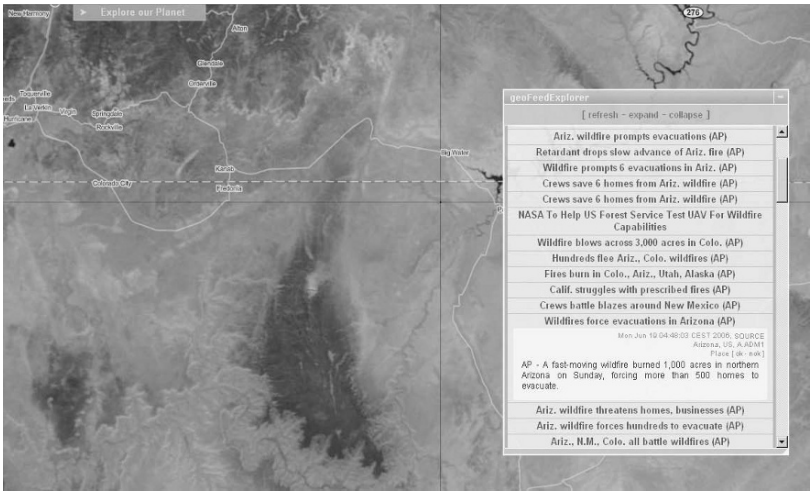


Figure 10.3: Map example with wildfire and corresponding RSS entry

The geoFeedExplorer is designed to handle missing elements, so that the interface does not look cluttered. The hierarchy is represented as a tree list, which responds to mouse wheel and scrollbar actions in a movable and resizable window and supports collapsing and expanding for all nodes. All feeds have three states: unloaded, loaded and expanded. Three main operations are possible in any state of the geoFeedExplorer:

- Refresh (loads or reloads the content of all registered feeds),
- Expand (shows all titles and map icons from all feeds, refreshes if needed),
- Collapse (hides all titles and map icons).

These operations are also available per feed and per title by simply clicking on an item. Only one click is needed after loading to display all entries and icons on the map. Icons and entries are linked; using a map icon will position the geoFeedExplorer to the desired entry, and a click on the *explore* icon will position and zoom the map. This gives the user a unique control and experience over the map and the content of the feeds. In the context of map and news items, it is clearly shown which country or region announces most news and builds up a hot spot.

The ability to configure their own list of RSS feeds subscriptions enables users to follow their own interests. The architecture is open to any topics whether climate, economy, open jobs or points-of-sale and displays perfectly the connection of news messages and the places they belong to.

10.4 The Big Picture

10.4.1 Feed Readers and Aggregators

RSS aggregators help users to keep track of read and unread feeds and feed entries. They offer tagging functionality and order feeds in a directory like a tree structure. But they offer very limited help or no help at all in grouping, filtering or prioritizing new entries. Users may get lost with dozens or even hundreds of feeds and an inbox overflown with unread items. Important and interesting entries are thus buried in an overabundance of feed entries since aggregators do not really help in finding relevant information.

Recommendation systems like findory.com help in finding new and interesting news items by analyzing users' click behavior. However, they do not analyze the content, they suffer from sparse data and they have a tendency towards recommending mainstream news.

Automatic information extraction with natural language processing may enable the next generation of RSS aggregators, which will help users manage information and automatically group, filter and prioritize news and blogs.

10.4.2 Tagging, Social Bookmarking and the Semantic Web

In the heydays of the first Internet boom, now referred to as Web 1.0, portals and compiled directories were the information hub of the Web. Human editors for commercial directory projects like Yahoo! or for free directories like dmoz,[38] the Open Directory Project, were classifying millions of Web sites in categories.

The newest generation of Web applications coined Web 2.0 has brought a renaissance of human-powered classification with social bookmarking and tagging. In contrast to one-dimensional directory categorization, tagging is inherently multidimensional and a Web page may be classified with an unlimited number of tags. A second important difference to the directory approach is the increase of *editors* by orders of magnitude. Now millions of users are tagging Web resources. Despite the huge popularity of tagging with early adapters, it is unlikely that the majority of average Internet users will actively tag things; consequently, most Web pages will remain untagged.

Similar to tags, publishers in the semantic Web encode meaning directly into the resource and make information accessible for automated processing. With better availability of free geodata and better geosupport in publishing tools, publishers will directly add semantic geographic information. However, the process has drawbacks and shortcomings. Publishers may refrain from this complex and error-prone task,

or it will also be exploited by spammers intentionally adding wrong semantic encodings to their site. Better-performing computer hardware components combined with free available geodata will increase the efficiency of automatic extraction of geo information from natural language text. Natural language processing may decrease the importance of manual tagging in the same way search engines have made directories obsolete.

10.5 Conclusions

The extraction of geospatial information from natural language text is a nascent technology, and we believe future accuracy improvements will be driven by mainly two factors. Faster computer chips with more processing power, on the one hand, will facilitate deeper grammatical text analysis. On the other hand, we expect better availability and accuracy of geographic data as GPS units are becoming common place and will help gathering and verifying geographic data. We also hope more governmental agencies will follow the United States' lead and release geographic data to the public domain.

Information extraction makes text accessible to further processing and enables its visualization on geobrowsers such as ExploreOurPla.net. The myriads of image layers on ExploreOurPla.net allow users to visualize news in its geospatial context and near-real-time satellite images. The fire and smoke of a wildfire are displayed as users read the news of the fire.

Our solution, which involves at least four parties, is an object example for Internet-based interaction of loosely coupled independent systems. The user (first party) is reading on the platform of ExploreOurPla.net (second party) content transparently enriched with geotags by geonames.org (third party). The author or publisher (fourth party) is freed from dealing with the technical complexity of adding semantic information and unwieldy geographical information systems.

Chapter 11

A Supervised Machine Learning Approach to Toponym Disambiguation

You-Heng Hu • Linlin Ge

Abstract*. This chapter presents a toponym disambiguation approach based on supervised machine learning. The proposed approach uses a simple hierarchical geographic relationship model to describe geographic entities and geographic relationships among them. The disambiguation procedure begins with the identification of toponyms in documents by applying and extending the state-of-the-art named entity recognition technologies and then performs disambiguation as a supervised classification processes over a feature space of geographic relationships. A geographic knowledge base is modeled and constructed to support the whole disambiguation procedure. System performance is evaluated on a document collection consisting of 15,194 local Australian news articles. The experiment results show that the disambiguation accuracy ranges from 73.55 to 85.38 percent depending on the running parameters and the learning strategies used.*

11.1 Introduction

Ambiguities exist in assigning a physical place to a given toponym in open domain unstructured and self-structured documents, such as free text and Web pages written in HyperText Markup Language (HTML). This is a natural consequence of the fact that many distinct places on the earth may use the same name. As a very simple example, the mention "Sydney" in "Sydney is a great holiday destination for families" could refer to many possible places around the world, such as Sydney, Australia; Sydney, Nova Scotia, Canada; or Sydney, Florida, United States. Real instances of such ambiguity can be easily found all around the world. Statistical data of the Getty Thesaurus of Geographic Names (TGN), a widely used geographic gazetteer developed by the Getty Information Institute, have shown that the percentage of toponyms that are used by more than one place ranges from 16.6 percent for Europe to 57.1 percent for North and Central America (Smith and Crane 2001).

The ambiguity of toponyms must be resolved to gain a full understanding of the geographic context of documents. A common and unique feature is that a toponym always refers to a place on the earth and thus corresponds to same geographic properties (e.g., geometry, topology and thematic data). This important feature makes the problem of toponym disambiguation different from disambiguation of general word sense and disambiguation of other proper nouns (e.g., personal names and organization names) in natural language processing.

Toponym disambiguation is one of the most important challenges in many geoinformatic applications including geographic information systems, geographic information retrieval and geographic digital libraries (Garbin and Mani 2005; Larson 1996; Zong et al. 2005).

In general, a toponym disambiguation procedure consists of two steps. First, a set of toponyms $T(t_1, t_2, t_3, \ldots, t_n)$ is extracted from a document. This step can be performed using software tools such as the Named Entity Recognition (NER) system. Second, the extracted toponyms and their context (i.e., the document from which the toponyms are extracted) are input into a toponym disambiguation algorithm, which then outputs a set of unique places $P(p_1, p_2, p_3, \ldots, p_n)$, each of which is the disambiguated referent of a toponym in the input set. The places in the output set can be expressed using numerical or descriptive formats; a geographic coordinate (e.g., latitude, longitude) is an example of the former, and a hierarchical administration structure (e.g., Australia/New South Wales/Sydney) is an example of the latter.

Two types of evidences can be used to support the disambiguation procedure: textual evidence and geographic evidence. For a toponym that is to be disambiguated, textual evidence refers to its linguistic environment, including linguistic syntax and discourse semantic, and geographic evidence refers to its related geographic properties, including geographic attributes (e.g., location and distance) and geographic relationships (e.g., administrative hierarchy and adjacency). Textual evidence can be discovered from local context using natural language processing techniques. On the other hand, external resources, such as geographic knowledge base and geographic gazetteers, are required to obtain geographic evidence. Both textual and geographic evidence play very important roles in the task. The different ways how these evidences are utilized and are integrated lead to different implementations of toponym disambiguation systems.

This chapter addresses the problem of toponym disambiguation by proposing a supervised machine learning approach. In particular, our approach consists of the following four components: (1) a geographic relationship model that is used to describe geographic entities and geographic relationships among them; (2) a geographic named entity recognition module that is used to extract geographic entities from context by applying the current state-of-the-art NER techniques; (3) a learning module that aims to build a disambiguation model from statistics of geographic relationships among geographic entities derived from a given training data set; the learned model then can be applied to new data sets that are similar to the training data;p and (4) a geographic knowledge base that provides the necessary underlying data resource to support the extraction and disambiguation procedures.

An experimental system that makes use of our approach has been implemented and evaluated on a large document collection that consists of 15,194 articles collected from the Australian Broadcasting Corporation (ABC) local news collection. The overall evaluation results provide quantitative support for the major hypothesis of this study that by employing the geographic relationship model and a general supervised machine learning algorithm, it is possible to improve the performance of toponym disambiguation.

The main contributions of this chapter are three-fold: first, the proposed approach is described, implemented and validated; second, a large document collection that can be used as a test-bed in future research is constructed; and last, a geographic knowledge base that contains hierarchical information about common place names of Australia is developed.

This paper is organized as follows. Section 11.2 provides background information related to the topic. Section 11.3 describes our methodology. Section 11.4 presents the experiments with the implementation of the algorithm, the obtained results and the important observations. Section 11.5 concludes this paper and gives some directions for future research.

11.2 Background

This section reviews some background information related to this work, including toponym disambiguation, geo/non-geo ambiguity and named entity recognition (NER).

11.2.1 Toponym Disambiguation

Various approaches have been proposed for the toponym disambiguation problem. Generally, these approaches can be grouped into three categories: corpus-based statistical methods, rule-based linguistics analysis and geographic heuristic approaches.

The corpus-based statistical methods, such as those proposed by Garbin and Mani (2005) and Smith and Mann (2003), apply statistical analysis and machine learning algorithms to a corpus in which toponyms are previously tagged and disambiguated in order to learn statistical models and classifiers that can be used to disambiguate toponyms in unseen documents. The key to the success of these methods is a good training set (i.e., the annotated corpus). However, the efforts and costs required to build a high-quality, broad coverage and large-scale annotated corpus that can be used as training data for toponym disambiguation are very significant (Leidner 2004).

The rule-based linguistics analysis approaches disambiguate toponyms based on their local context using linguistic pattern-matching and co-occurrence analysis (Rauch et al. 2003; Zong et al. 2005). Examples of some widely used heuristic in these methods include "a toponym is qualified by its following toponym" and "toponyms in one document share the some geographic context". By applying these rules, the true reference of the toponym "Sydney" in, for instance, "Many people think Sydney, New South Wales, is the capital of Australia" can be easily found. Rule-based linguistics methods are simple and easy to implement. However, for better disambiguation performance, it is useful to combine them with other evidences such as default sense and population data (Amitay et al. 2004).

The geographic heuristic approaches (Smith and Crane 2001; Woodru and Plaunt 1994) perform geographic computations for disambiguation based on geographic nature of each possible candidate place. Possible geographic features that can be used include geographic distance and area size. Rules like higher scores are assigned to those candidates that are geographically closer to other places that have no ambiguity and to those candidates that are bigger in area are used as heuristic guidance to find the right candidate. The major issue with these methods is that an external knowledge base must be integrated to provide necessary geographic data.

11.2.2 Geo/Non-Geo Ambiguity

Beside the ambiguity of toponym in the geographic domain, there is another type of ambiguity called geo/non-geo ambiguity (Amitay et al. 2004). Geo/non-geo ambiguity happens in the following cases.

First, a toponym is used as its nongeographic synonyms. For examples, in English some surnames are taken from place names, such as York and Lancashire (Jobling 2001). On the other hand, many places in the world are named after people. Examples in Australia include Darwin, the capital city of the Northern Territory, which was named after Charles Darwin, the British naturalist. Additional, some places were named using common words. Sunshine and Waterfall are two examples that can be found in Australia.

Second, a toponym is used as a metonymy that refers to another related concept, such as a governmental and community body. As an example from the ABC News (August 30, 2004) indicates, "Australia", the country name in the context of "Australia won medals in 14 different sports in Athens with the golds spread over six sports", refers to its sport team.

Third, a toponym is included in attributive phrases, such as "the chef from China" and the "Prime Minister of Australia". In these examples toponyms (i.e., China and Australia) are used as an attribute to their relational nouns and not necessarily to refer to any geographic location.

11.2.3 Named Entity Recognition

Named entity recognition (NER) is a research domain that aims to identify and classify instances of different types of named entities from text. The Message Understanding Conference standard defines seven named entity types for NER tasks, i.e., <person>, <organization>, <location>, <time>, <date>, <monetary value> and <percent> (Chinchor 1997).

Existing NER techniques can be categorized into three groups: those that use linguistic analysis such as Humphreys et al. (1998), those that use machine learning techniques such as Yangarber and Grishman (1998), and those that combine the two previous approaches, such as Mikheev et al. (1998). To perform NER tasks, the first approach applies grammar rules and gazetteers that are developed by experienced linguists to process text, and the second approach applies statistical models that are learned from large amount of training data.

NER is a very important component of many natural language processing (NLP) applications. In this work, NER techniques are used to recognize place names in text.

11.2.4 Geographic Relationship Model

Geographic relationships between geographic objects reflect the nature of their embedding in the real world. The geographic relationship model in our approach is designed to describe geographic objects to which toponyms are mapped and the geographic relationships that hold between geographic objects. Specifically, the model defines a simple hierarchical structure that is used to map toponyms to geographic objects, and three geographic relationships that are considered in the disambiguation procedure.

Our approach maps toponyms to a hierarchical structure, which is based on the political administrative properties of toponyms. Nodes on the higher level of the hierarchy consist of one or more nodes on the lower levels in a political sense. A three-level country/state/city hierarchy is an instance of the structure. By using this structure, a toponym can be mapped to a single place in the world.

The reasons that this mapping strategy is selected are (1) it is simple and well understood; (2) compared with other possible mapping strategies that are concerned with geometric properties (e.g., location coordinates, boundary edge and bounding boxes), this approach requires less computation and storage cost; and (3) there are many existing resources available for building a geographic knowledge base that uses this structure as internal data structure.

Based on the above hierarchical structure, three alternative relationships that are possible between two geographic objects are defined in our geographic relationship model, namely: *identical*, *similar* and *part-of*.

Identical: Two geographic objects are identical if they have the same values at all levels of the hierarchical structure. Same as the "one sense per discourse" heuristic in NLP literature (Gale et al., 1992), in many cases, several mentions of a toponym in a document are mapped to a signal geographic object. However, different toponyms may refer to identical geographic objects as well, since the same place can have multiple names. For example, both *Stalingrad* and *Volgograd* are mapped to the same famous Russia city on the west bank of Volga River. This issue also has an important impact on multilingual applications.

Similar: Two geographic objects are similar if (1) they are at the same hierarchical level, and (2) they have the same upper level values or both of them are the topmost level objects. For example, *Australia/New South Wales* and *Australia/Queensland* are similar, because both of them are states of Australia. *Australia* and *China* are similar, because both of them are country-level geographic objects.

Part-of: A geographic object *X* is part-of another geographic object *Y* if *X* is at a lower hierarchical level than *Y*. For example, *Australia/New South Wales/Sydney* is part of *Australia/New South Wales*, *Australia/New South Wales* is part of *Australia*.

The part-of relationship in the hierarchical structure is different from the part-of concept derived from the geographic geometry point of view, where the former focus on the political relationship between two geographic objects, but the latter checks whether a geographic object is entirely surrounded by another one. Enclaves and exclaves are real-world examples that describe this difference. These relationships are helpful to resolve toponym ambiguities. Let's consider a simple example.

Example 1: Many people think Sydney is the capital of Australia.

Two toponyms can be extracted from the above sentence: *Sydney* and *Australia*. The toponym needed to be disambiguated is *Sydney,* which can be mapped to many geographic objects, including

- Australia/New South Wales/Sydney,
- Canada/ Nova Scotia/Sydney,
- United States/North Dakota/Sydney,
- United States/Florida/Sydney.

Based on the geographic relationships between *Sydney and Australia,* it is easy to find out the correct mapping (i.e., *Australia/New South Wales/Sydney*), because there is a part-of relationship between this mapping and *Australia,* and no relationship can be found using other mappings. Now, let's look at an interesting question arising from the following examples.

Example 2: China, Texas, is located in Jefferson County.
Example 3: In the United States, Texas is the third largest exporter to China.

It is easy for a human reader to understand the context and then realize that the toponym *China* is mapped to two different geographic objects in the above two examples: *United States/Texas/China* for the former, and *China* for the latter.

The ambiguity of *China* in Example 2 can be resolved easily, as there is only one relationship (i.e., *United States/Texas/China* is part of *United States/Texas*) can be found from the context for the toponym *China.* However, for *Example 3*, the ambiguity of *China* is more complex because many possible relationships can be found:

- United States/Texas/China is partof United States/Texas;
- China is similar to United States.

The disambiguation procedure for this example in our approach is regarded as a classification problem in which each toponym has ambiguities classified into one of its possible geographic object mapping. Relationships connected to geographic objects are used as the classification features, based on which ambiguities are resolved by using supervised machine learning technique.

11.2.5 Geographic Named Entity Recognition

The goal of the geographic named entity recognition module is to identify all toponyms in a given document. This procedure is carried out by employing the state-of-art NER technologies based on gazetteer-based string matching and statistical analysis methods. Two important extensions are also made to improve the system performance.

The first extension is called "geographic stop words". In information retrieval, stop words are words that do not have semantic meaning (e.g., a, the) or occurs in many of the documents in the collection (e.g., say, you). These words are not useful for information retrieval tasks and can be eliminated during the information process procedure in order to reduce the processing and storage costs. The concept of the geographic stop words in our approach is similar, but with a focus on geographic place names, and only applies to the gazetteer-based string matching method. Examples of toponyms that can be seen as geographic stop words include US, Mobile (a city in Alabama), Orange (a city in Texas, a city in New South Wales, Australia) and Reading (a town in Berkshire, U.K.). The existence of these words in a document introduces geo/nongeo ambiguity to the whole toponym disambiguation procedure when the gazetteer-based string matching method is used. By applying a geographic stop word list, this kind of geo/nongeo ambiguity could be efficiently removed and the size of the NER result sets could be reduced. Therefore, the number of toponyms that are needed disambiguated could be reduced, and the computational costs of further processing could be reduced as well.

The second extension is called "multi-words toponym merging". Many toponyms consist of more than one word, and in some cases, a part of a toponym could be a toponym as well. Examples include New South Wales (a state of Australia) in which both "South Wales" and "Wales" could be used as toponyms that refer to an area of England. The multi-words toponym merging algorithm finds a maximum overlap to merge two or more geographic named entities into one toponym by checking their positions in the documents.

In summary, the whole geographic named entity recognition procedure can be described as a procedure of three steps. The first step performs a simple string matching against all documents in the collections utilizing the gazetteer derived from our geographic knowledge base, and toponyms in the geographic stop word list are eliminated during this step. The second step performs an NER process using a pre-trained NER tagger to tag three types of named entities: <person>, <location> and <organization> in all documents. The final step matches result sets from the two previous steps using following rules: (1) for each string that found in the first step, it is eliminated if it is tagged as a non-location entity (i.e., <person> or <organization>) in the second step, otherwise it was added to the result set; (2) for each toponym in the geographic stop word list of the first step, it is added to the result set if it is tagged as a <location> entity in the second step; (3) two or more toponyms are merged using the above multi-words toponym merging algorithm.

11.2.6 Supervised Machine Learning

After all toponyms are identified and tagged in a document, a supervised machine learning module is used to disambiguate all toponyms that could be mapped to more than one geographic places. The disambiguation procedure can be treated as a supervised classification procedure, and the statistics of geographic relationships among toponyms are used as features for classification.

Typically, four steps are involved in a classification process based on supervised machine learning methods, such as Naive Bayes classifier, logistic regression and decision trees.

First, one or more quantitative characters are selected as features for representing data entities to be classified. The collection of these features is usually referred to a feature space, which is normally modeled as a multidimensional vector space. The selection of the features is problem dependent and can be done both manually and automatically.

Second, a statistical model is chosen based on background knowledge and domain knowledge. By using this statistical model, the total feature space is divided into a number of subspaces, each of which corresponds to a class. Both linear and nonlinear functions can be used to describe this model.

Third, a classification rule is constructed for the statistical model and is used to estimate the class-conditional probabilities (i.e., the likelihood that a given data entity has a particular set of feature values) and the probability of appearance of each class for any data entity to be classified based on Bayes' rule of conditional probability, in which prior probabilities are estimated from training data.

And last, the class probabilities calculated in the previous step are used to predict a class for each data entity to be classified based on the maximum likelihood estimation, i.e., the class that has the maximum probability of the observed feature values is selected as the result class for a given data entity.

Our proposed classification method covers all the above-mentioned steps. The feature scheme used is based on statistics of geographic relationships among possible mappings of toponyms. Here we discuss the detail of the classification scheme, feature selection and feature value acquisition.

Let T $\{t_1, t_2, t_3, \ldots, t_n\}$ be the set of all toponyms that is extracted from a given document, and let P_k $\{p_{k1}, p_{k2}, \ldots, p_{km}\}$ be the set of all possible geographic place mapping of t_k. The problem of disambiguation of t_k is defined as a classification procedure that assigns one element of P_k $(p_{k1}, p_{k2}, \ldots, p_{km})$ to t_k. As an example, considering the above Example 1, we have T = {*Sydney, Australia}* and P_{Sydney}= {*Australia/New South Wales/Sydney*, *Canada/Nova Scotia/Sydney*, *United States/North Dakota/Sydney*, *United States/Florida/Sydney*} and $P_{Australia}$ = {*Australia}*.

Once all toponyms are identified and all their possible mappings are acquired, a weighting value that is used as the quantitative feature for classification is assigned to each element in all P sets. The weighting algorithm is described as follows:

Algorithm 1 GeoRelationshipWeighting

```
(1)    procedure GeoRelationshipWeighting
(2)    input: T
(3)    output: weighting_matrix (t, p)
(4)
(5)    for each t of T
(6)      for each p of Pt
(7)        weight = 0.0
```

```
(8)          for each t' of T
(9)            if t == t' then continue
(10)           for each p' of Pt'
(11)            if (p identical to p') then
(12)              weight = weight + IDENTICAL_WEIGHT/size(Pt')
(13)            else if (p is part of p') or (p' is part of p) then
(14)              weight = weight + PARTOF_WEIGHT/size(Pt')
(15)            else if (p is similar to p') then
(16)              weight = weight + SIMILAR_WEIGHT/size(Pt')
(17)            end if
(18)          end for
(19)        end for
(20)        weighting_matrix (t, p) = weight
(21)      end for
(22)      end for
```

The constant values (i.e., IDENTICAL_WEIGHT, PARTOF_WEIGHT and SIMILAR_WEIGHT) in the GeoRelationshipWeighting algorithm are decided by experiments.

11.2.7 Geographic Knowledge Base

The geographic knowledge base is our approach provides a rich repository from which all necessary information for geographic named entity recognition and geographic relationship analysis can be acquired. The data schema of our geographic knowledge base is defined using the object-oriented modeling method. Figure 11.1 shows the class diagram of the schema, in which four classes are defined: (1) a *geographic entity* class that is used to describe a distinct place on the earth. Properties of the *geographic entity* class include *id, qualified name* and *type* (e.g., city, state, and country); (2) a *geoname* class that is used to represent names of geographic entities; (3) a *relationship* class that is used to model the geographic relationships between geographic entities; and (4) a *part-of* class, which is a subclass of the *relationship* class.

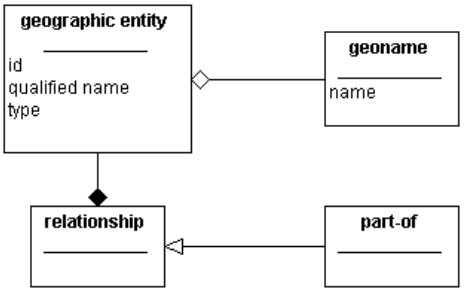


Figure 11.1: Class diagram of the geographic knowledge base data model

Figure 11.2 gives an example of the instance-level view of the model. A relational database management system (RDBMS) is used to implement the geographic knowledge base.

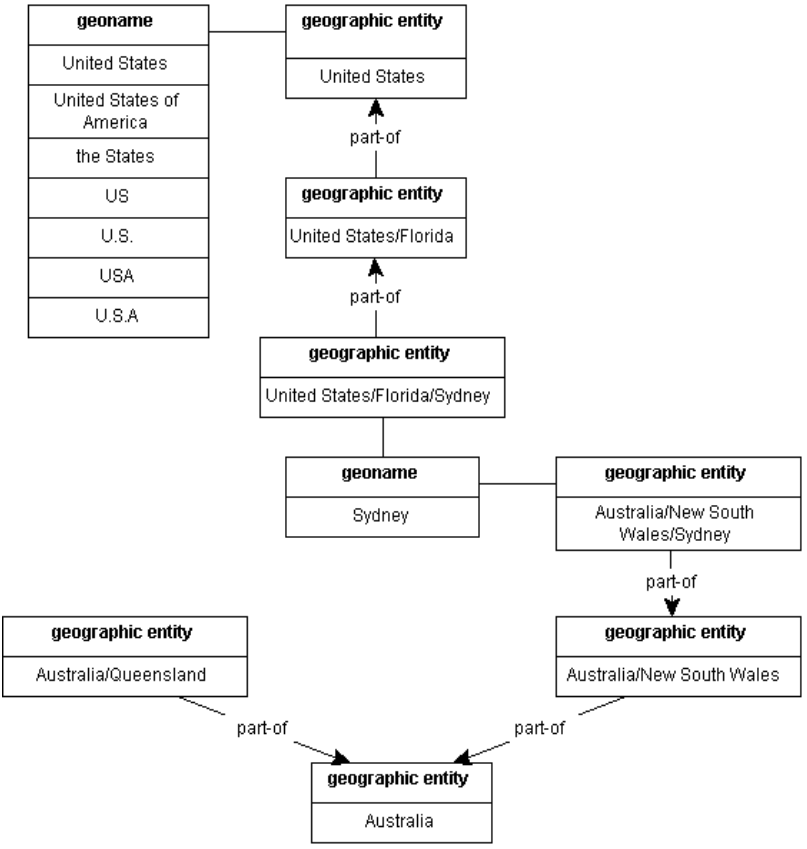


Figure 11.2: An instance-level view of the geographic knowledge base data model

11.3 Experiments

An experimental system has been implemented to evaluate the proposed method. This section first describes the document collection used in the experiments and then presents baseline methods and the evaluation of our system.

11.3.1 Data

Our experimental system aims to disambiguate toponyms of Australia. The toponym disambiguation process is modeled as a procedure that classifies each toponym that has ambiguities into a set of eight state-level geographic entities (i.e., New South Wales, Queensland, South Australia, Tasmania, Victoria, Western Australia, Northern Territory and Australian Capital Territory).

A geographic knowledge base instance is constructed based on the data model presented in Section 11.2.7. Statistics from the knowledge base are summarized in Table 11.1. The two resources from which geographic knowledge is acquired are the Gazetteer of Australia, developed by Geoscience Australia, and the Postcode Datafile provided by the Australia Post.

Table 11.1: Statistics from the geographic knowledge base for Australia toponyms disambiguation

State/Territory	Total Number of Distinct Toponyms	Number of Ambiguous Toponyms	Percentage of Ambiguous Toponyms
New South Wales	2,454	274	11.17
Queensland	2,396	249	10.39
South Australia	826	132	15.98
Tasmania	605	121	20.00
Victoria	1,692	245	14.48
Western Australia	1,681	167	9.93
Northern Territory	112	10	8.93
Australian Capital Territory	123	30	24.39
Australia	9,203	1,228	13.34

The document collection used in our experiments consists of 15,194 articles collected from the Australian Broadcasting Corporation (ABC) local news collection. The number of unique toponyms found in this collection is 1,179, and the number of toponyms having ambiguities is 762. The ground truth of all toponyms in the collection is obtained through subjective viewing.

11.3.2 Evaluation

To evaluate the performance of our system, experiments were carried out using 10-fold cross validations over the ABC local news collection, and the results are averaged over all 10 runs. The two baseline methods used for comparison are:

Baseline 1 – Maximum occurrence: Given a toponym to be disambiguated, this method counts all candidates (i.e., state-level geographic entities) from the training data set, and returns the candidate with the maximum occurrence. If the toponym is not part of the training data, a default mapping determined by experts is assigned.

Baseline 2 – Maximum local weighting score: This method runs the GeoRelationshipWeighting algorithm to calculate weighting scores for all toponyms and their mappings in a document. The mapping with the maximum weighting score is assigned to a toponym. A default mapping is used when the maximum weighting score is zero (= no geographic relationship can be found for a toponym).

Three supervised machine learning algorithms are trained and applied: Naive Bayes classifier, J48 decision trees and boosted J48. These algorithms are implemented by using the Weka software package developed at the Department of Computer Science, University of Waikato, New Zealand.

Three runs were performed, each of which used different parameter configurations as shown in Table 11.2.

Table 11.2: Parameter configurations used in the experiments

	Run 1	Run 2	Run 3
IDENTICAL_WEIGHT	1.0	1.0	1.0
PART OF_WEIGHT	1.0	0.8	0.5
SIMILAR_WEIGHT	1.0	0.8	0.5

Table 11.3: Toponyms disambiguation accuracy (%) on the ABC local news collection

	Run 1	Run 2	Run 3
Baseline 1	78.71	78.71	78.71
Baseline 2	51.85	52.33	52.11
Naive Bayes classifier	75.62	73.70	73.55
J48	85.37	84.70	84.58
Boosted J48	85.38	85.08	84.73

Table 11.3 shows the experiments' results. In all three runs, the J48 and boosted J48 algorithms performed consistently better than the two baseline methods; the Naive Bayes classifier performed better than baseline 2, but worse than baseline 1. The best accuracy result was 85.38 percent using the boosted J48 in run 1.

The best results for the three machine learning algorithms (i.e., 75.62 percent for Naive Bayes classifier, 85.37 percent for J48 and 85.38 percent for boosted J48) were all found in run 1, configured as IDENTICAL_WEIGHT = 1.0, PARTOF_WEIGHT = 1.0 and SIMILAR_WEIGHT = 1.0, and their accuracy results decreased when the values of PARTOF_WEIGHT and SIMILAR_WEIGHT decreased.

11.3.3 Discussion

From the above results several observations emerge. First, the statistics of toponyms of Australia confirm the claim that toponym ambiguity is a common fact in real-world practice. The percentage of toponyms with ambiguities ranges from 8.93 percent for Northern Territory to 24.39 percent for Australian Capital Territory, and an overall value of 13.34 percent is found for the whole Australia. These figures show that resolution of toponym ambiguities must be taken into account in any geographic-related natural language processing and information retrieval applications.

Second, the results from the above experiments clearly show that the overall performance in disambiguation accuracy can be significantly improved by employing the proposed method and appropriate supervised machine learning algorithms. This observation indicates that our approach including the geographic relationship model and the geographic knowledge base is a very promising one to be applied to the toponym disambiguation problem.

Last, we are aware that the performance (e.g., the recall and precision measures) of the geographic named entity recognition module has not been fully evaluated. The main reason for this limitation is that we lack necessary human and technical resources to annotate large document collections. However, the effectiveness of the module could be tested with existing annotated corpus and golden standard data. Further experiments are planned to address this issue.

11.4 Conclusions and Future Work

This chapter presents a toponym disambiguation approach based on supervised machine learning. In particular, the details of four components, including a geographic relationship model, a geographic named entity recognition module, a learning module and a geographic knowledge base, have been discussed. The proposed approach has been evaluated over a large collection of local Australian news articles. The experiment results demonstrate that our algorithms, in particular, those using the J48 and boosted J48 machine learning methods, can provide better performance than other baseline methods.

Future work is planned in two directions. The first is to extend the current implementation to validate the proposed approach by evaluating the performance of worldwide toponym disambiguation tasks with a focus on acquiring geographic knowledge. The second is to discover and utilize other types of information entities that may be useful to improve disambiguation performance. Examples include telephone numbers, postal codes, famous local people's names and local events mentioned in context.

Chapter 12

Geospatial Information Integration for Science Activity Planning at the Mars Desert Research Station

Daniel C. Berrios • Maarten Sierhuis • Richard M. Keller

Abstract. *NASA's Mobile Agents project leads coordinated planetary exploration simulations at the Mars Desert Research Station. Through ScienceOrganizer, a Web-based tool for organizing and providing contextual information for scientific data sets, remote teams of scientists access and annotate data sets, images, documents and other forms of scientific information, applying predefined semantic links and metadata using a Web browser. We designed and developed an experimental geographic information server that integrates remotely sensed images of scientific activity areas with information regarding activity plans, actors and data that had been characterized semantically using ScienceOrganizer. The server automatically obtains remotely sensed photographs of geographic survey sites at various resolutions and combines these images with scientific survey data to generate "context maps" illustrating the paths of survey actors and the sequence and types of data collected during simulated surface "extra-vehicular activities." The remotely located scientific team found the context maps were extremely valuable for achieving and conveying activity plan consensus.*

12.1 Introduction

Through the proliferation of high-speed communication networks and wide-spread availability of desktop computing systems, researchers in many different fields should now be able to conduct data gathering and analysis campaigns that involve larger groups of collaborating scientists separated by great distances. However, the design requirements for computer-based systems to support these efforts are not yet fully known. There is some evidence that functions such as synchronous electronic chat and data annotation capabilities (Olson et al. 1998) and geospatial displays for locating and planning scientific data collection (Ogren et al. 2004) can be quite useful. But roles for many other functions remain to be established or elucidated. For example, the relationship between the nature of data collected (e.g., qualitative vs. quantitative), collection and analysis methods employed (e.g., automated vs. nonautomated), or the domain of investigation and the optimal design of systems supporting scientific collaboration is still unclear. In this study, we discuss our experience developing and deploying a system for generating geospatial and temporal traces of scientific data through dynamic integration of semantically tagged information.

The Mars Desert Research Station (MDRS) near Hanksville, Utah, in the United States is one of two experimental Mars habitats developed and built by the Mars Society (marssociety.org); the other habitat is located on Devon Island in the high Arctic. For several years, the MDRS has been the location of planetary exploration

simulations conducted by NASA's Mobile Agents project (Clancey et al. 2002). Part of the research activities conducted during these simulations includes studies of the process of distributed scientific collaboration, activity planning, sharing and review of collected scientific data, and the design and development of computer-based tools to support these processes. We have had the opportunity to participate in these investigations through our work developing and deploying ScienceOrganizer (Keller et al. 2004) during Mobile Agents MDRS fields tests in 2003 through 2005.

ScienceOrganizer is a Web tool for managing contextual information (Dey et al. 2001) of scientific data sets (Berrios et al. 2004) specifically developed to support the work of distributed, collaborating scientific and engineering teams. Through ScienceOrganizer, remotely located teams of scientists can access and annotate data sets, images, documents and other forms of scientific information, supplying additional metadata and interconnecting them through predefined logical relationships using any Web browser. Information stored thusly in ScienceOrganizer is semantically characterized along multiple dimensions, providing users with more precise "tags" with which to find data compared to traditional information storage systems. In addition, this semantic tagging can provide users with valuable cues regarding information purpose, provenance and pedigree and can assist in information navigation. Finally, semantic tags can be used to support two functions of advanced scientific information systems: information integration and inference.

During 2005, we specifically sought to study how temporal and geospatial metadata can be used to organize and present scientific information for improved access and sharing. In this chapter, we discuss the development of an experimental geographic information server (GIS) that integrated remotely sensed images of scientific activity areas (field sampling sites, geographic features, etc.) with information regarding activity plans, actors and data that had been characterized semantically by software agents and stored in ScienceOrganizer. The server automatically obtained remotely sensed photographs of geographic survey sites at various resolutions and combined these images with scientific survey data to generate maps illustrating the geospatial paths of survey actors and the temporal sequence and types of data collected during simulated surface "extra-vehicular activities" (EVA). This integration played key roles in support of scientific decision making for activity planning and execution prior to and during EVA.

12.2 Methods

The primary purpose of the GIS is to support the needs of remotely located scientists to help plan and monitor the activities of a geological exploration team in the field. It leverages the semantic metadata stored in ScienceOrganizer to formulate requests for geographic, topographic and photographic information from a publicly available archive. After obtaining this information through Web services, the GIS synthesizes temporal and geospatial maps showing the precise sequence and location of scientific activities and data products collected during simulated EVA.

12.3 Mobile Agents Architecture

NASA Ames' Mobile Agents Architecture is a distributed agent-based software architecture that integrates diverse mobile entities in a wide-area wireless system for simulated lunar and planetary surface exploration (Clancey et al. 2005). Software agents, implemented in the Brahms multi-agent language, run in virtual machines onboard laptops integrated into space suits or robots or located in habitats (Clancey

et al. 1998; Sierhuis et al. 2002). "Personal agents" support the crew in the habitat and on the surface, who communicate with the agents via a speech dialog system. All the actors (human and robotic agents) in the simulations are outfitted with high-precision global positioning devices that continuously track their locations. Data (e.g., digital images, voice recordings, sample measurements, etc.) are collected during EVA simulations according to a science activity plan generated by a collaborating team of remotely located scientists. Software agents transmit the collected data via a dynamically configured wireless network to an installation of ScienceOrganizer located in the habitat. These agents generate and tag the data with a predefined set of metadata that varies depending on the type of data collected. However, for collected scientific data, these metadata always include the GPS location of the agent that collected the data. The crew in the habitat and remotely located scientists can then view real-time data or metadata in ScienceOrganizer (see Figure 12.1).

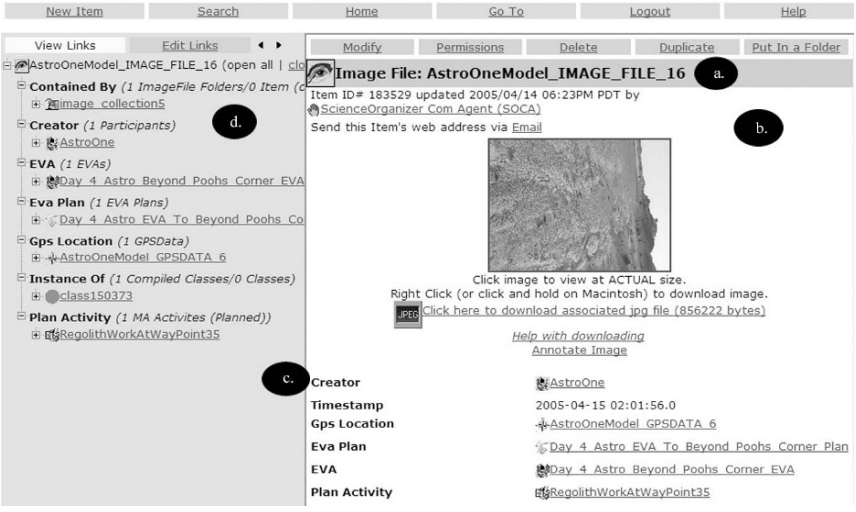


Figure 12.1: The Web application, *ScienceOrganizer;* this display shows details of one data item, an "Image File", including its name (a), a "thumbnail" version of the image (b), and a portion of its metadata (c) and semantic links (c, d) to other data items

The participants in the 2004 and 2005 field tests collected many images of sampling sites and surrounding areas and recorded voice notes describing major land features. While these data can provide a context for current and past activities, and help plan for future activities, it has proved difficult for remotely located participants (in a "surface" habitat or in a mission control center) to relate data products to other data products or to activities temporally and geospatially. With advances in the remote sensing capabilities of satellites orbiting earth, and the now-widespread availability of the high-resolution images, we sought to develop a system for dynamically integrating these images with the (metadata) collected at MDRS.

12.3.1 TerraServer-USA

TerraServer-USA is a Web service/site that provides access to remotely sensed aerial photographs and topographic maps of the earth's surface (limited to the United States; terraserver.microsoft.com) for a range of spatial resolutions. TerraServer-

USA offers a number of Web services in addition to image access, including area/region identification and gazetteer functions. Image resolution varies by geographic area, with very high resolution imagery (the kind required for supporting scientific activities such as precise instrument deployment or manipulation) available only in limited (mostly urban) areas. Images are offered as fixed-size tiles (each identified by a unique set of metadata) from which client programs can compose larger images ad hoc.

12.3.2 Selecting and Integrating Information

The GIS uses the semantic data and metadata generated by the ScienceOrganizer Communication Agent (a Brahms agent) and attached to various kinds of scientific data when it stores them in the ScienceOrganizer information system (Figure 12.1). These metadata provide key contextual information for collected scientific data. For example, as a new survey panoramic image is stored in ScienceOrganizer, its "EVA" property is set to a reference to the particular EVA simulation during which it was collected, its "GPS data point" property is set to a reference to the latitude/longitude coordinates where the photo was taken, etc. The GIS gathers this contextual knowledge from ScienceOrganizer and then generates requests for imagery from TerraServer-USA using this information. It then combines the returned imagery using the contextual information of the region of the EVA to yield a context map image (Figure 12.2).

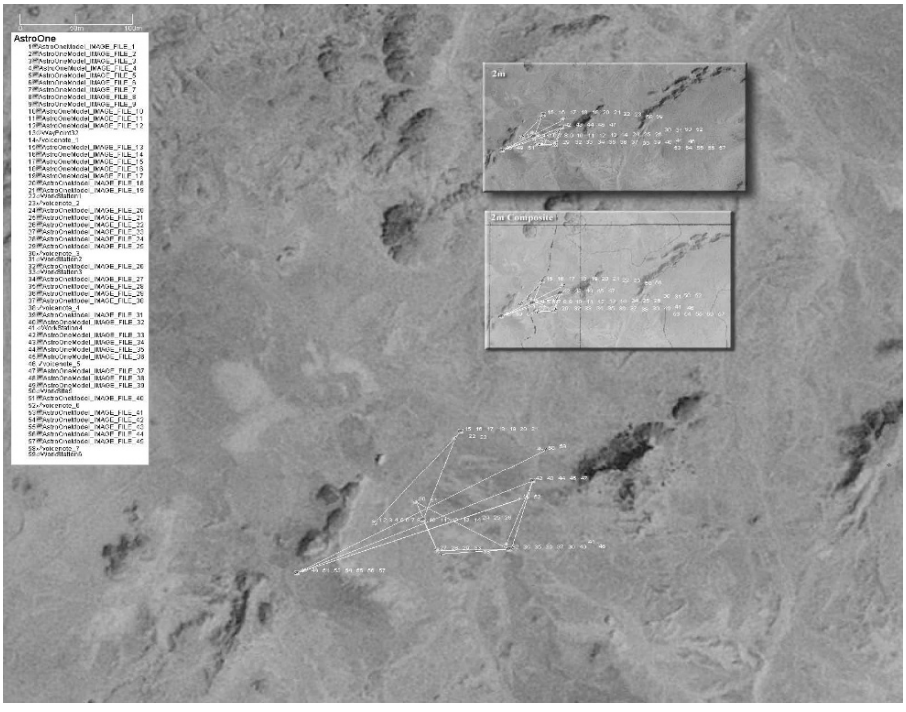


Figure 12.2: Context map image created by the GIS and stored in ScienceOrganizer; the larger image is a 1-m-resolution photograph (overlaid EVA information); the smaller inset images are a 2-m-resolution photographic map and a 2-m-resolution composite photo/topographic map with the same overlay

The contextual information used by the GIS to generate context maps includes the scientific area-of-interest bounding box, inferred from the GPS coordinates of all the scientific data products collected during the EVA simulation (as of the time of map generation). The bounding box (plus additional area for displaying a map legend and scale) determines the size of final context map image. The GIS obtains the necessary tile images from TerraServer-USA to span the entire bounding box and then stitches these images together to form the base layer of the context map image. Next, small icons are placed and sequential numbers drawn at the locations of each of the data products according to their order of collection. Data products collected closely in time and space often can result in overlapping and unreadable numbers; in such cases, we chose to displace subsequent numbers horizontally (from left to right, as in Figure 12.3). Finally, points of collection are connected by (straight) lines to indicate rough traverse paths of agents. Future versions of the GIS could produce maps with more precise (nonlinear) path traces, as actor GPS coordinates are ascertained every second.

Because actors involved in the simulations frequently cross paths, we programmed the GIS to use knowledge of the collecting actor to sort data and create a context map for each actor (Figure 12.3). This minimizes confusion between actor traces on the final map; however, these actor-specific map images do not show shifting spatial relations between actors over time, which may be important for certain collaborative scientific activities (e.g., collection tasks that require two or more actors working with a specified distance between them). Also, there are clearly some situations in which showing all path traces on a single map would prove valuable (e.g., for robot path planning). Finally, we repeated the map image creation process using aerial photographic images at spatial resolutions of 1 m and 2 m, using topographic maps at 2-m resolution and, finally, compositing 2-m topographic and photographic images with semi-transparency (as shown in Figure 12.2, inset). We also transformed 1-m photographic images by doubling their dimensions to simulate 0.5-m context map images. These maps proved valuable in displaying simulations with tightly spaced activities, in spite of the reduced clarity of map images.

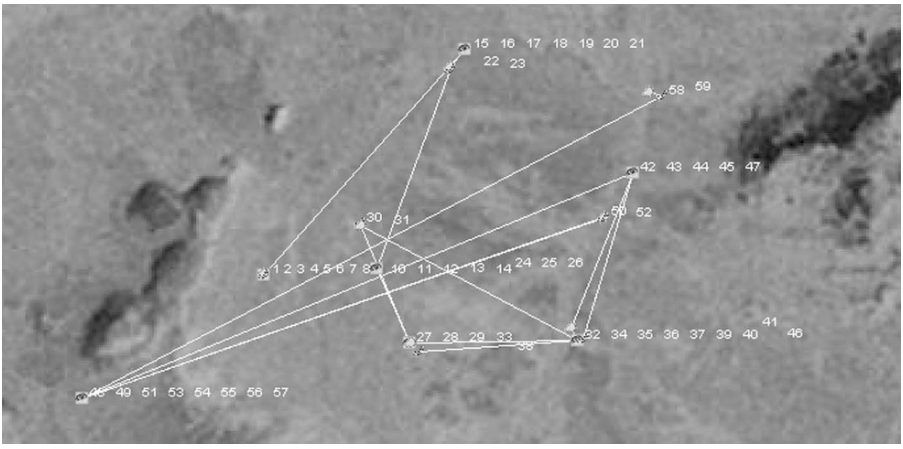


Figure 12.3: A 1-m-resolution GIS-generated map showing the activities of a single actor, "Astro-One", during a simulation conducted April 14, 2005; "Rock Hill" is recognizable at left; collected data items are numbered sequentially (beginning due east of "Rock Hill"); note evidence of imperfect determination of collection sequence (e.g., nos. 32, 33 and 34), due to erroneous timestamps of large data files

In addition to constructing these context map images, the GIS also generated HTML image maps and combined them with the context map images to form "dynamic" (user-actionable) area maps that it also stored in ScienceOrganizer. Each data product shown on these maps was linked to its (unique) location (URL) in ScienceOrganizer so that users could readily move from context map image to hyperlinked data product (see Figure 12.4).

Finally, the GIS stored all context map images and map documents in ScienceOrganizer as soon as they were generated, generating semantic links between the maps and the simulations shown on the maps. This created an efficient way for users to navigate from map to simulation to data products and back. As the EVA simulation progressed, the GIS updated these map images and documents, enlarging the EVA area bounding box (and map) size, and adding additional data collection points, actor paths and links to data products in ScienceOrganizer continuously. If new actors joined the simulation, new maps and documents displaying their collected data locations were created. Using these maps, remotely located ScienceOrganizer users were able to follow the activities of the field scientists in near real time.

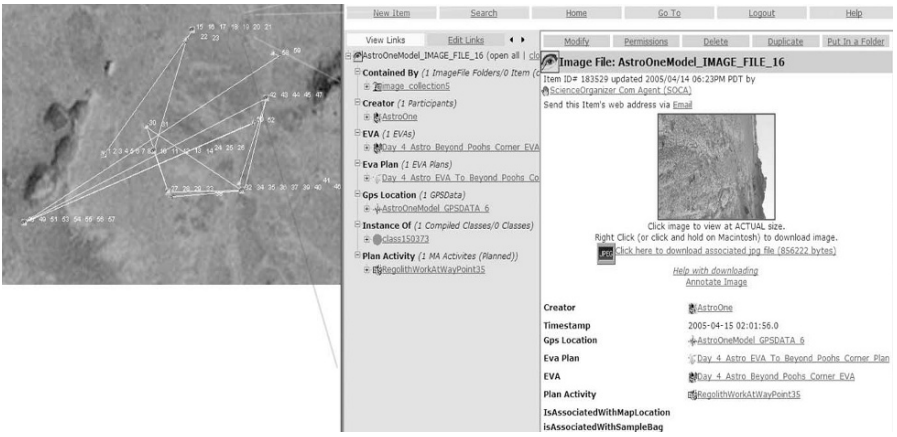


Figure 12.4: Use of an image map to link to a data product collected during the EVA shown in Figure 12.2

In Figure 12.4, the representation of a collected digital photograph (a type of "Image File") is selected by clicking on the ordinal number "19" on the image map (i.e., the 19th collected data product), after which the user is shown the item in ScienceOrganizer (right, including a thumbnail of the photograph, its metadata and semantic links to other contextual information in ScienceOrganizer).

12.4 Results: MDRS 2005

We had the opportunity to test the performance of the GIS and the use of the context maps in EVA simulations conducted at MDRS in April 2005. Overall, generation of the map products was considered timely by the participants, although there was often a delay of one or two minutes from data collection to map generation. The GIS was programmed to search continuously for active simulations and then create context maps so that users could track the progress of scientific activities in near real time, overwriting any previous context maps (since the data on those maps were no longer "current"). For the kinds of geological survey and sampling work performed at MDRS and the use of the maps by remotely located scientists, the de-

lay in map generation was not deleterious. Periodically, the Web services offered by TerraServer-USA became unavailable. This resulted in map products being deleted but not replaced by the GIS with updated versions. Users complained that maps had mysteriously evaporated on a couple of occasions. About halfway through the field tests, we built into the GIS an adaptive capability to detect the availability of TerraServer-USA and abort map generation before deleting existing maps when the service was unavailable.

Having fine-tuned the performance of GIS, we then sought evidence of whether the context maps produced were useful to the remotely located team of 10 scientists who were guiding the field geologists, specifically whether they aided them in planning the field team's next set of activities. The first author participated in the Remote Science Team (RST) planning meetings held before selected simulations during the field tests. These meetings were held by voice teleconference and using a desktop/whiteboard sharing application (WebEx.com). Prior to the field tests, one of the field geologists had independently obtained aerial photographic maps of the sites likely to be surveyed and had annotated the maps with ad hoc feature names (see Figure 12.5). The RST adopted these feature names and frequently used them to identify data products in planning-meeting discussions (e.g., "sample 3 taken at Red Hill"). The context maps proved extremely useful for coordinating agreed-upon features with specific data products; the team found it very difficult to distinguish and refer to data products based solely on collected GPS coordinates. Furthermore, the RST found that the context maps were extremely valuable for conveying consensus plans for the following day of simulation activities. In particular, they used the image-annotation capabilities in ScienceOrganizer (see Figure 12.6) to draw these plans on the context maps themselves, indicating which areas in the region the field scientists should investigate next with geometric shapes (circles, ovals, etc.) and words (e.g., "Sample here").

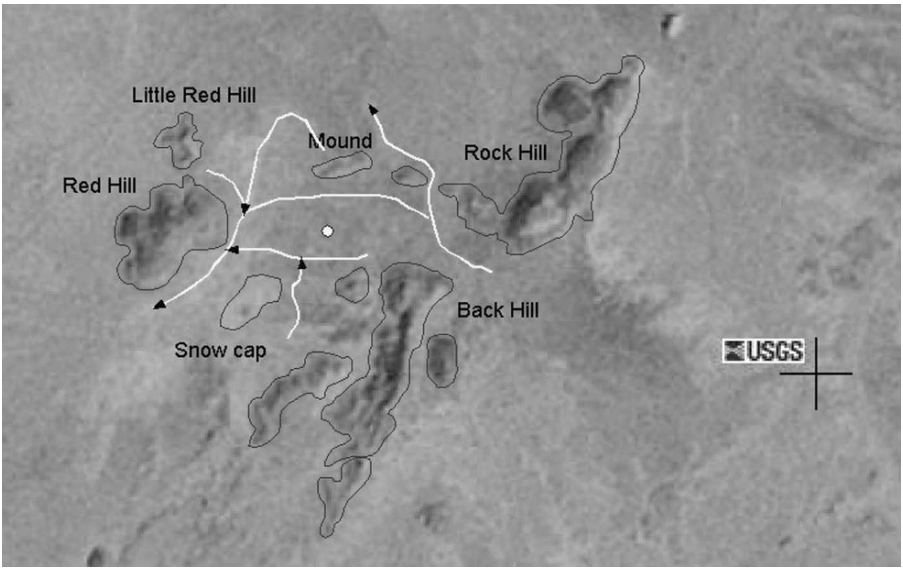


Figure 12.5: Annotations of composite USGS images of proposed geologic survey sites at MDRS drawn manually by a field geologist; black outlines indicate major land features, and white lines indicate likely fluvial traces

The Brahms agent responsible for storing collected data (images, voice notes, etc.) in ScienceOrganizer was designed to tag those data with (time of collection) timestamps only after the agent had completely received the data. This design yielded occasional errors in the sequence of data products as shown on the context maps (see Figure 12.3), because larger data files required significantly longer transmission times over the wireless network in the field. However, we found no evidence that this kind of relatively small imprecision in temporal sequencing significantly impacted planning or other science activities during the simulations.

One feature that the context maps lack is reference to a coordinate system. The field team requested that the RST deliver GPS waypoints that could be incorporated into specific activity plans required by the Mobile Agents Architecture. A grid overlaying the maps showing latitude and longitude tick marks would have been very useful for this purpose.

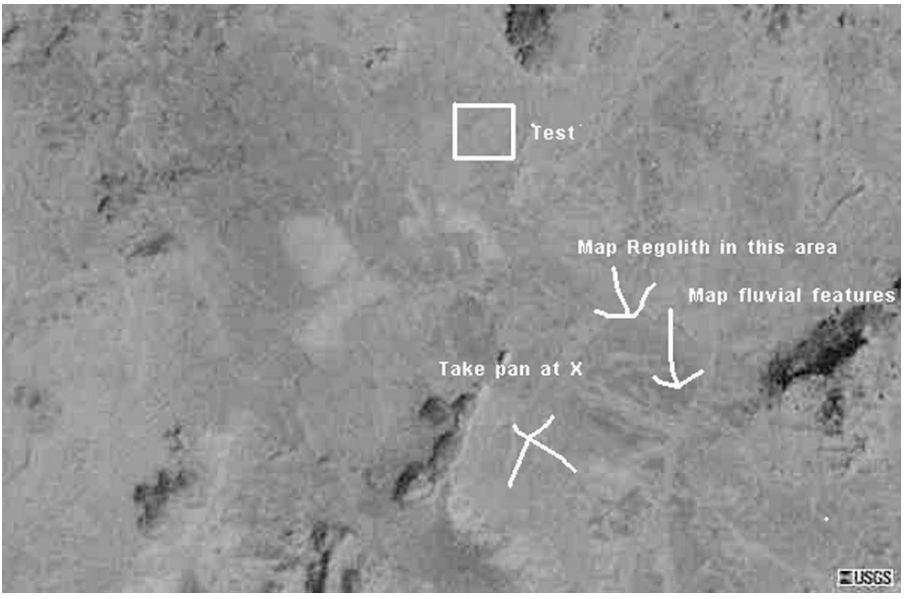


Figure 12.6: A human-annotated map image of the same area shown in Figures 12.2 and 12.3

After examining the GIS map shown in Figure 12.2, one of the remote science team members uploaded a cropped image (obtained through USGS) from her laptop to ScienceOrganizer and then used the ScienceOrganizer Image-annotator to draw notional instructions for the field geologists directly on the image (the box at the top represents an initial test by the scientist of the Image-annotator's capabilities). Instructions included (1) "take pan (i.e., panoramic photograph) at (location) X", (2) "map regolith in this area" (indicated by the arrow) and (3) map fluvial features" (in the area indicated by the arrow).

12.5 Related Work

The use of "collaboratory" applications like ScienceOrganizer in which multiple actors share data and contextual knowledge is growing (Olson et al. 1998; Bebout et al., 2002; Chin and Lansing 2004) but is still far from widespread. This has limited

the ability to study the nature and utility of various types of geospatial contextual information; certainly the optimal way to integrate geospatial, temporal and other types of information to trace moving objects that have arbitrary paths and relationships remains to be demonstrated. There is some knowledge as to the most efficient methods to conflate geospatial information from heterogeneous sources in order to show locations of stationary objects [e.g., buildings (Michalowski et al. 2004)] and to locate moving objects with predefined paths [e.g., trains (Shahabi et al. 2001)], although the actual usefulness of this information in real-world settings that involve collaborative decision making has yet to be shown.

Planetary surface visualization applications and frameworks that can provide sophisticated, three-dimensional views of terrain with information overlays from other sources – e.g., Google Earth[2] and NASA World Wind[1] – could be used to provide the same type of geospatial context cues as the GIS. For the type of collaborative work involved in the Mobile Agents project, this approach presents some important issues to explore. Can the use of these sophisticated tools be incorporated into the workflow of teams like Mobile Agents as successfully as the GIS? What is the value of the type of dynamic geospatial views produced by these applications, compared with the static but persistent kind of context maps generated by the GIS?

12.6 Discussion

Coordinating medium- and large-scale scientific data gathering campaigns presents many challenges, especially when some actively involved participants are not physically located at or near collection sites. We focused on supporting the collaboration between such participants and the rest of the team with computer-based tools that allow all participants to monitor campaign progress, view scientific data gathered with coordinated contextual geospatial, temporal and geographic information, and jointly plan on-going campaign activities. We had the opportunity to test the utility of tools we developed as part of an existing software agent system to simulate the collaboration of in situ and remotely located scientists cataloguing sites of interest and collecting scientific samples. This kind of distributed collaboration is becoming the model for more and more scientific data gathering campaigns, and closely parallels the current (late-mission) collaboration model followed by participants in the on-going Mars Exploration Rover missions (Wick et al. 2005).

The remotely located science team made significant use of the technologies we developed to guide humans and robots during scientific field investigations. Our primary role during these experiments is to support such collaboration but also includes attempting to identify obstacles to conducting such investigations efficiently. The participants in the collaboration were very receptive to trying new computer-based tools for communication and planning, and this must be kept in mind when comparing our observations to others in this field.

We observed scientists' use of context map images and image maps to discuss and guide sample collection in the field, arguably the most important function of remote scientific teams participating in such campaigns. During the 2005 field tests that we observed, the members of the remote science team found the maps, which depicted the precise sites of all samples collected, as well as locations where images and voice notes were recorded, very valuable for coordinating agreement on recommended plans for further field investigation. This contrasts with our observations in 2004 of similar field studies by the same team that did not have the type of maps produced by the GIS at their disposal. During the 2004 field tests, the scientists frequently spent a great deal of time trying to determine temporal sequence of

collected data, to associate data with locations and to grasp and communicate the geospatial relationships of these locations during planning meetings. The type of data product access and visualization provided by the context image maps during the 2005 field tests reduced the frequency of such "down-time" activities, leaving the scientists more time to concentrate on discussing concerns relevant to domain scientific work.

Certainly there are many aspects of the map products we have presented that could be further optimized through more careful, controlled study and evaluation. For example, the layout of data product numbers at each collection site was merely the most obvious solution, and with additional resources we could have explored more advanced techniques using layered or three-dimensional visualizations of data product numbers. We look forward to continuing our evaluation of the usefulness of our work through future exploration simulations at MDRS.

Acknowledgements. We wish to acknowledge the significant contributions of Dr. Bill Clancey and Ron Van Hoof, Mike Scott, Ian Sturken, David Hall, Matt Linton and Shawn Wolfe. We also acknowledge our field geologists, Abigail Semple and Dr. Brent Garry, and all members of the Mobile Agents Remote Science Team led by Prof. Shannon Rupert.

Chapter 13

Inferences of Social and Spatial Communities over the World Wide Web

Pragya Agarwal • Roderic Béra • Christophe Claramunt

Abstract. *The research presented in this chapter introduces a graph-based and computational modeling approach to the analysis of Web-based networks. The aim is to derive a social network and compute its emerging spatial and thematic properties from the semantics embedded in a series of Web pages. We apply several graph-based operators and complement them with thematic, spatial and similarity operators. The principles of the modeling approach are applied to the study of research communities as they appear over the World Wide Web. This allows us to infer the degree of correlation between the different properties of the semantic networks that emerge from research communities on the World Wide Web.*

13.1 Introduction

The Semantic Web, where computers are empowered with machine-processable semantics, provides promising and effective search of information resources (Berners-Lee et al. 2001). The research presented in this chapter explores the role of the Web in the development of social and community networks and applies computational and graph-based methods to analyze and visualize the emerging network properties. These properties are analyzed with respect to the geographical, thematic and semantic dimensions.

A social network is commonly defined as a set of people who share a common interest and have connections of some kind (Wasserman and Faust 1994). Social networks provide useful insights into ways that the social communities are formed and interact. They have been widely studied over the past years, particularly in mathematical and statistical research (Strogatz 2001; Newman 2003), and have also been applied in many application domains such as epidemiology (Moore and Newman 2000), environment (Dunne et al. 2002) and scientific citation (Redner 1998). Nowadays, the World Wide Web provides many opportunities for inferring and analyzing social networks, as Web sites often embed information with respect to a specific domain of interest and the actors that are part of it (Berners-Lee et al. 2001; Greco et al. 2002; Hou and Zhang 2002; Bekkerman and McCallum 2005). The research presented in this chapter introduces a social and spatial network approach to infer and analyze the structure and properties of a semantic network derived from the relationships embedded and hidden in a subset of the Web. Our objective is to complement graph-based operators with spatial, thematic and network correlation measures that cannot be easily expressed by a graph layout but can be virtually derived from the relationships embedded and hidden on the Web. This should favor the understanding of the role of space and thematic properties in a given social network. We are particularly curious to study the degree to which network properties are correlated to properties in the geographical dimension.

The domain used for our study is the semantic network formed by the scientific community closely related to the field of geographical information science (GIS). Network and graph-based analyzes support exploration of patterns of collaborations in this field and reveal to what degree trajectories of researchers impact the formation of a network of communities. To illustrate the approach we consider two major international conferences of the field: the Conference on Spatial Information Theory (COSIT) and the Conference on Geographical Information Science (GIScience), which present the advantage of reflecting the key players of the GIScience community. These two conferences appear relatively central in the field of GIS and are well documented on the Web. The objective of the study is not to study individuals and their networks, but rather to analyze the affiliation network of the universities and research centers that have supported the careers and progression of these researchers. The expected output is a network of universities active in this particular research field. We also assume that the trajectory of a given researcher from one university to another represents an implicit connection between these two universities. This information is derived and computed from the short biographies as they appear on the Web pages of the researchers hyperlinked to the conference Web sites. This supports derivation of a semantic network whose properties are analyzed using graph measures and similarity measures on the one hand, and thematic and geographical-based correlations on the other hand.

This chapter extends a series of works related to the analysis of spatial networks when applied to urban systems (Béra and Claramunt 2003) and in Web environments (Agarwal et al. 2006). The work presented here develops further the notion of similarity when assessing the commonalities between different semantic networks. The remainder of the paper is organized as follows. Section 13.2 introduces the principles of the modeling and computational approach. The case study is developed in Section 13.3 and the main results presented. Finally, Section 13.4 draws some conclusions and outlines future work.

13.2 Semantic Network Modeling

A semantic network is modeled as a hypergraph $G(N, E)$, where N is a finite set of nodes and E is a finite set of links between these nodes. A node is denoted as n_i, a link between two nodes $n_i, n_j \in N$ as a pair (n_i, n_j).

Although network analysis usually combines local and global measures, global measures mainly outline the emergent structure of the represented graph, while local-based measures characterize particularities in the graph. Global measures evaluate the role of a node in the graph. These include, for example, the average distance between nodes, the ratio of shortest paths a node lies on (Sabidussi 1966), and the accessibility of a node (Batty 2004; Béra and Claramunt 2003; White and Smyth 2003). A common measure of centrality is the *betweenness centrality* $C_b(i)$ that gives the fraction of shortest paths between node pairs that pass through a given node (Freeman 1977; Brandes 2001) and is defined as

$$C_b(i) = \sum_{k \neq j \neq i} \frac{S_{jk}(i)}{S_{jk}}, \tag{1}$$

where S_{jk} denotes the number of shortest paths from j to k and $S_{jk}(i)$ is the number of shortest paths from j to k that i lies on. The *betweenness centrality* can be generalized to multiple component graphs, as well as for directed graphs.

Although these graph-based measures are expected to exhibit the global properties of a given network, they do not take into account the spatial and thematic dimensions. This leads us to introduce some additional geographically related measures that correlate the structural and geographical properties of the semantic network. From a given semantic network $G(N, E)$, we consider the geographically based subsets of N that group the nodes of N according to their membership to a given classification and partition of the underlying space (the same principles can be applied to a thematic classification and a measure of thematic dispersion). The objective is to analyze to which degree the structure of a network is correlated to a given classification. Let us consider a classification

$$C = \{c_1, c_2, \ldots, c_j, \ldots, c_p\}$$

that forms a partition of N with $p > 1$. The local measure of heterogeneity is defined as the ratio of the number of adjacent nodes of n_i belonging to another class to the number of adjacent nodes belonging to the same class as n_i (Agarwal et al. 2006):

$$Het(n_i) = \frac{|n_{i\rightarrow}|}{|n_{i\leftarrow}| + 1}. \tag{2}$$

A refinement is given by a weighted measure of heterogeneity Het(n_i) so as to give more importance to well-connected nodes, regardless of the homogeneity or heterogeneity of the nodes considered in terms of class membership. This weighted measure is given as follows (Agarwal et al. 2006):

$$WHet(n_i) = (|n_{i\rightarrow}| + |n_{i\leftarrow}|) \cdot Het(n_i) = \deg(n_i) \cdot \frac{|n_{i\rightarrow}|}{|n_{i\leftarrow}| + 1}, \tag{3}$$

where $\deg(n_i)$ is the degree of node n_i.

Measures of geographical heterogeneity can be defined at different levels of granularity depending on the level of abstraction chosen, and different partitions of the same geographic space can be considered as semantic networks are interrelated in different ways, particularly on the Web where they can appear implicitly or explicitly across different Web sites. The subjective component of the nature of these relationships means that a strategy is needed in order to evaluate the degree of semantic relationships and/or similarity between two given semantic networks. Given (and assuming) the fact that a semantic network can be qualified with some well-defined labels, an intuitive measure of similarity between two semantic networks is given by a cross comparison of the number of pages where they are present in one form or another, or in other words by comparing their respective footprints left on the Web.

More formally, let $\mathrm{h}(q)$ denote the set of hit pages returned by a search query q, q being a set of keywords $\{k_1, k_2, \ldots, k_n\}$ that qualify a given semantic network N. A similarity measure between two semantic networks N_1 and N_2, respectively, materialized by the search queries q_1 and q_2, can be defined by the number of hits returned by the combined query $q = q_1 \cup q_2 = \{k_{11}, k_{12}, \ldots, k_{1l}\} \cup \{k_{21}, k_{22}, \ldots, k_{2m}\}$.

This search query yields a set of hit pages, $h(q) = h(q_1 \cup q_2) = h(q_1) \cap h(q_2)$, since the pages fulfilling both queries q_1 and q_2 in the query $q_1 \cup q_2$ are at the intersection of pages fulfilling query q_1 and pages fulfilling query q_2. The number of hits retuned by the query $q_1 \cup q_2$ is $|h(q_1 \cup q_2)| = |h(q_1) \cap h(q_2)|$. The similarity measure between search query q_1 and search query q_2 is defined as (Agarwal et al. 2006)

$$Sim(q_1, q_2) = \frac{|h(q_1 \cup q_2)|}{|h(q_1)|}. \quad (4)$$

13.3 Research Community Analysis

The modeling approach is applied to the Web sites of the series of conferences GIScience 2004 and COSIT 2003 and 2005. Only full papers accepted to these conferences were considered: 26 for COSIT 2003, 31 for COSIT 2005, and 25 for GIScience, so that all three samples can be considered of comparable size. A Google search was performed from the (complete) author and university names, looking for CVs and résumés on personal and/or departmental Web pages, in order to develop, as accurately as possible, trajectories from academic career paths. Figure 13.1 illustrates how the mediation between a conference and a university, represented by the straight links, can be inferred from several Web-based resources. The conference's Web site, and the underlying information on paper authors, is used to make a link to an author's CV through his or her homepage. Additional information is derived when necessary from a publication database. Many personal homepages (or by default, the ones of the affiliated universities) include online CVs or similar biographic information including life trajectories along universities.

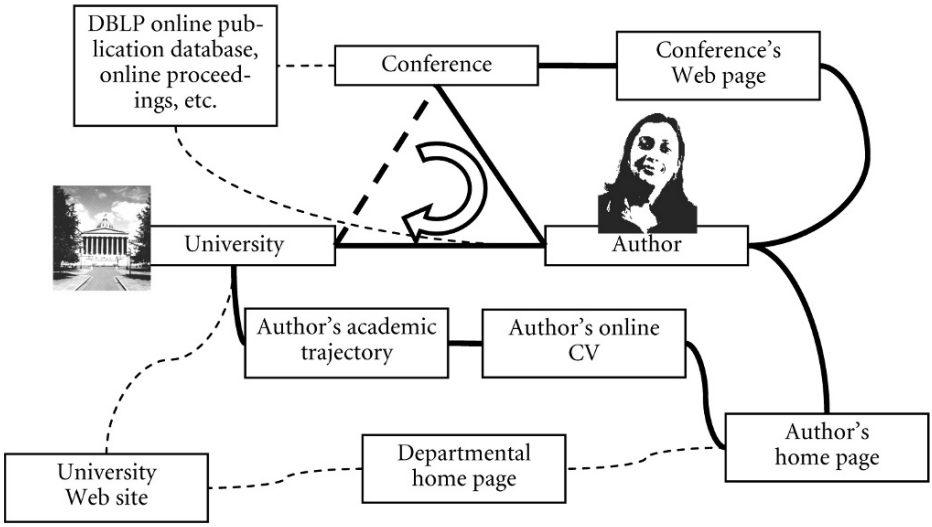


Figure 13.1: Link between a conference and a university mediated by an author (central triangle); curved links show the Web-based connections involved in this mediation, enabling the extraction of academic trajectories; dashed lines show secondary connections, providing extra knowledge on the network

Significant academic changes considered are any substantial changes in the university associations with a given researcher. We chose not to use any prior or offline knowledge, to ensure that the World Wide Web is the main source of information. We are aware that some of the trajectory information may be missing. This is the case for authors not maintaining a personal Web page or in departments providing little information for their staff. The online proceedings, accessible from Springer's Web site, and the online reference database of the University of Trier (DBLP) were also helpful in deriving this trajectory information.

The universities derived from the academic career paths of the authors of the papers presented at these conferences provided the elements for the identification of the semantic network nodes, along with the relationships between the universities inferred from the academic trajectories. For example, the event of an academic who got her PhD at the University College London (UCL) and then moved to the University of Maine followed by the University of Edinburgh gives two directed links, UCL–Maine and Maine–Edinburgh, but also between UCL and Edinburgh in order to reflect cumulative relationships. This leads to the appearance of clusters and networks of universities within the graph.

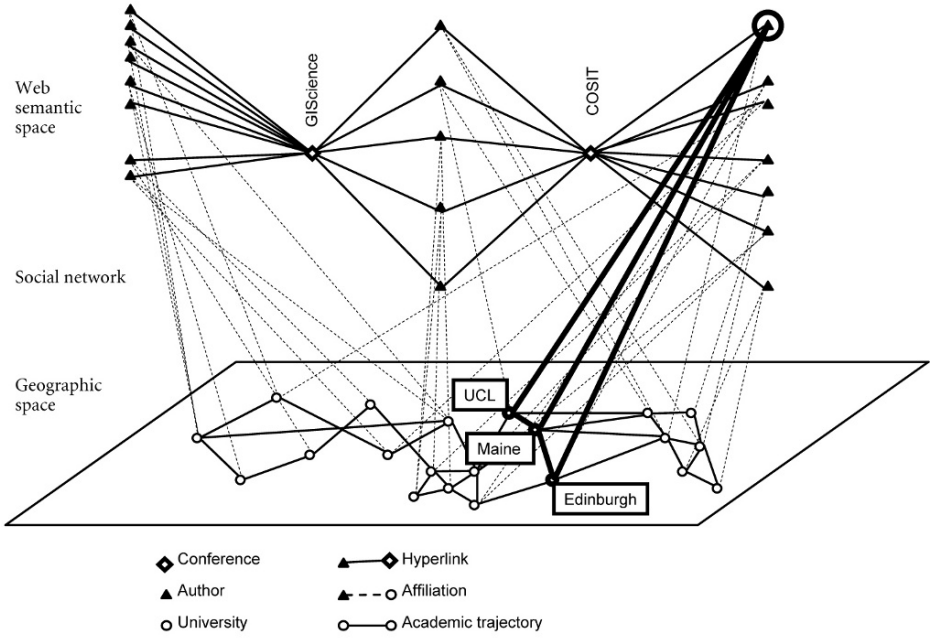


Figure 13.2: Semantic networks derived for conference authors from the Web indicative of social networks when superimposed over geographic locations; highlighted is trajectory example

It is expected that the analysis of the graph and geographically based emerging properties of the semantic network from the information implicitly available over the Web information space should help in making apparent, and for qualifying the degree of integration of, the research community. Figure 12.2 presents how the superimposition of the semantic network derived from the Web and the geographic space derived from the world map leads to the derivation of the social network. In this case, the footprints of the different trajectories are also made apparent.

13.3.1 Structural Analysis

The patterns and trends that we consider in our analysis are key university players in the research area of interest, connections between these universities, degrees of compactness versus spread of the research community, clusters and peripheries. The networks of universities derived from the two conference Web sites have several emerging characteristics as illustrated in Figure 13.2 (note that the authors are represented in a semantic space made of three overlapping domains: COSIT 2003, GIScience 2004 and COSIT 2005).

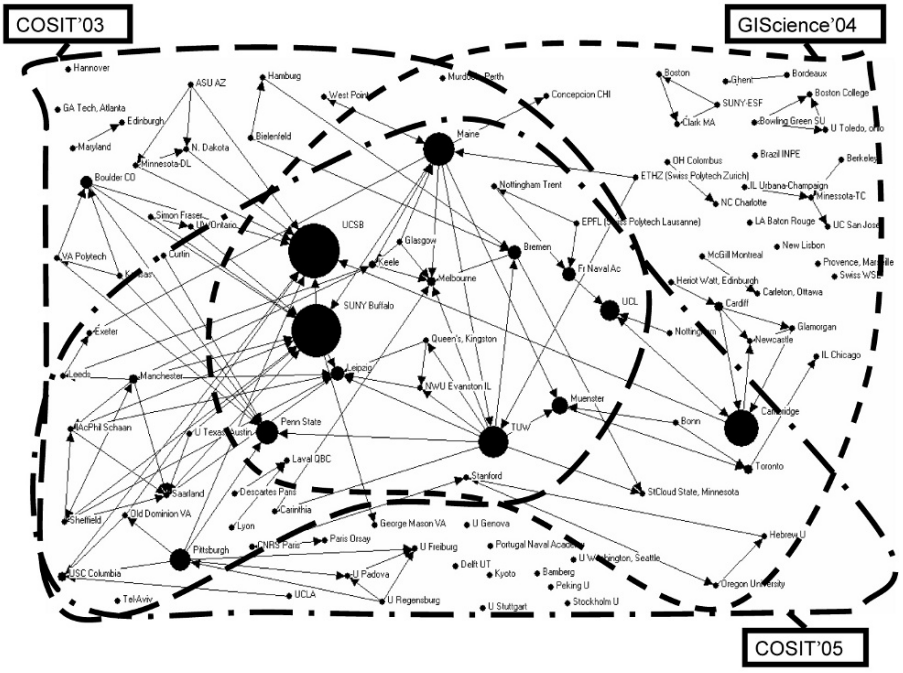


Figure 13.3: Network of universities derived from COSIT 2003, COSIT 2005, and GIScience 2004; the nature of involvement is materialized by three overlapping sets; the central area (intersection of the above-mentioned sets) is that of the universities present in all three events; the size of the nodes reflects the magnitudes of their betweenness centralities

Figure 13.3 shows in- and out-degrees through the number and direction of edges joining the network's vertices, and the *betweenness centrality* measure (undirected, with normalization of values belonging to different components) of nodes is represented by their size. This figure, while showing the primary networks for all the three events, also shows the common space they share, provided by the actors and networks that have contributed to all three conferences. The central component of the graph connects most of the universities, with highly connected and attractive nodes reflecting "key players", as represented by the relative sizes of the nodes in this case. Beside centrality, connectivity (in- and/or out-degrees) is a way of assessing the "popularity" or attractiveness vs. repulsiveness of a node. For instance, this is noticeable for the University of California at Santa Barbara (UCSB), the State University of New York, Buffalo (SUNY Buffalo) and the University of Melbourne. The high (betweenness) centrality value, amounting to the larger size of the node, is a

measure of the number of authors connected in one way or another during their trajectories to these key players. Other highly attractive nodes are related to the academic changes of a very mobile researcher, who connects six universities together (i.e., SUNY at Buffalo, the University of Leipzig, Saarland University, Schaan, Sheffield and Manchester). Also, some vertices are highly connected but appear rather repulsive (high out-degree, when compared to their in-degree), as is the case of the Technical University in Vienna (TUW). This is due to the fact that many junior researchers, who get their doctoral diploma at TUW, move to other sites in the network, providing an important "input" and attractiveness measure to other nodes in the research network. Only few other small isolates are present.

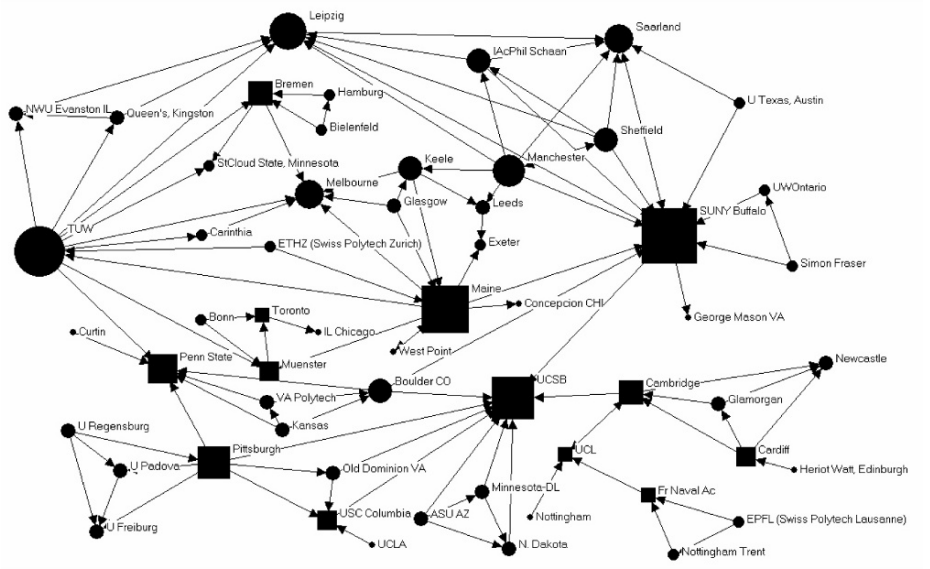


Figure 13.4: Illustrative network showing the role of "key players" in maintaining the connectivity of the graph's main component; cut points appear as squares; node sizes are proportional to their degrees of connectivity

As a measure of the integration (or "unavoidability") of nodes, the *betweenness centrality* measure is indicative of how likely the members of the research community are to move through these nodes, which tend to be also well connected (thus forming hubs and authorities). Taking connectivity into account, the higher values are obtained essentially for universities involved in all three events (COSIT 2003 and 2005, GIScience 2004), and for the ones that have most authors associated to them along their academic trajectories.

The semantic networks show that the connectivity characteristics of COSIT differ from those of GIScience. Whereas in COSIT the graph is strongly dominated by a main central component made of 33 nodes, GIScience, on the other hand, is less structured with respect to the academic trajectories, composed of numerous smaller and isolated components. Moreover, the community in GIScience conferences is less connected to the networks composing the COSIT conferences, and the nodes have relatively lower betweenness centralities, at least in parts of the network that are not common to both conference series. This suggests that the academic trajectories have played a smaller role in establishing the community for the GIScience con-

ference as compared to the one for COSIT. However, it seems clear that there is some other process at work in the emergence and cohesion of the GIScience network, and that this process has a strong geographic component since almost exclusively North American universities are involved. Also, it is worth noticing that in the central component of the graph, all of them except one (Concepcion University in Chile) are involved in COSIT 2003, whereas only a third of the total were involved in the GIScience conference.

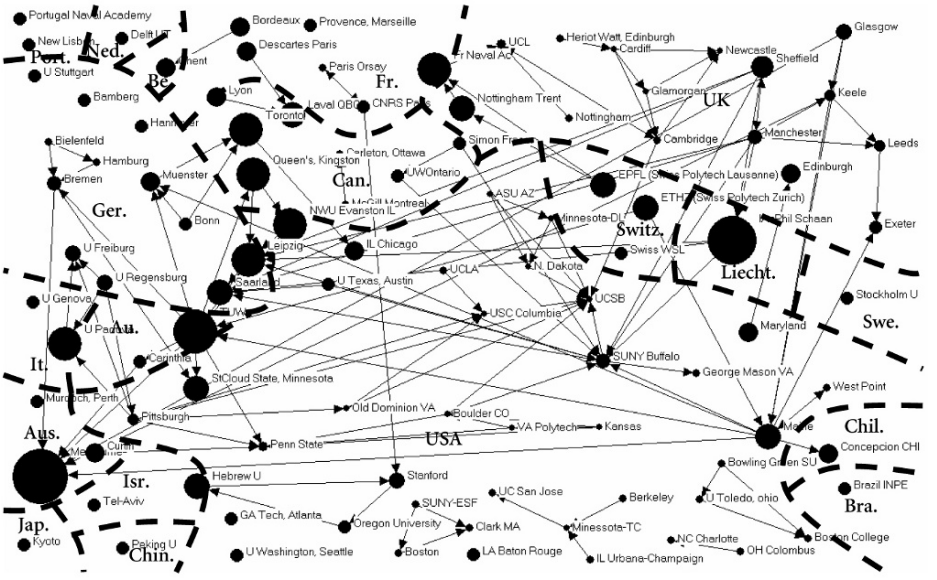


Figure 13.5: Unweighted intranational heterogeneity

The key players in the network can also be considered from another perspective showing clearly their central role in maintaining cohesion in the graph and their roles as a bridge between smaller communities that otherwise would have been disconnected. Figure 13.4 shows these key players as cut points as well as nodes of high degree within the graph's main component. Overall, high betweenness centrality, high degree and cut point situation appear to be correlated, as in the case of TUW, Maine, SUNY or UCSB.

13.3.2 Heterogeneity Analysis

National and continental networks and patterns can similarly be analyzed using the heterogeneity index [cf. Eq. (2)], which gives relative higher weighting to those nodes that are more connected to the nodes outside the host nation than to other nodes inside it. This also provides a comparative view on networks using different thematic parameters. In this case (see Figure 13.5), when applying the heterogeneity index to an intranational semantic space, it is noticed that University of Melbourne stands out as the only prominent node within Australia. On the other hand, several universities in Canada appear central and prominent. TUW maintains a central position, too, as related to its intranational network, with researchers migrating to universities in other countries. Similarly, the universities of Saarland and Leipzig emerge as central in the research network as they support connections and move-

ments of researchers to other nodes on the research network, more to countries lying outside Germany than within it. Within the UK, the universities fail to get a high magnitude of heterogeneity, since most trajectories are intranational, or at least intracontinental.

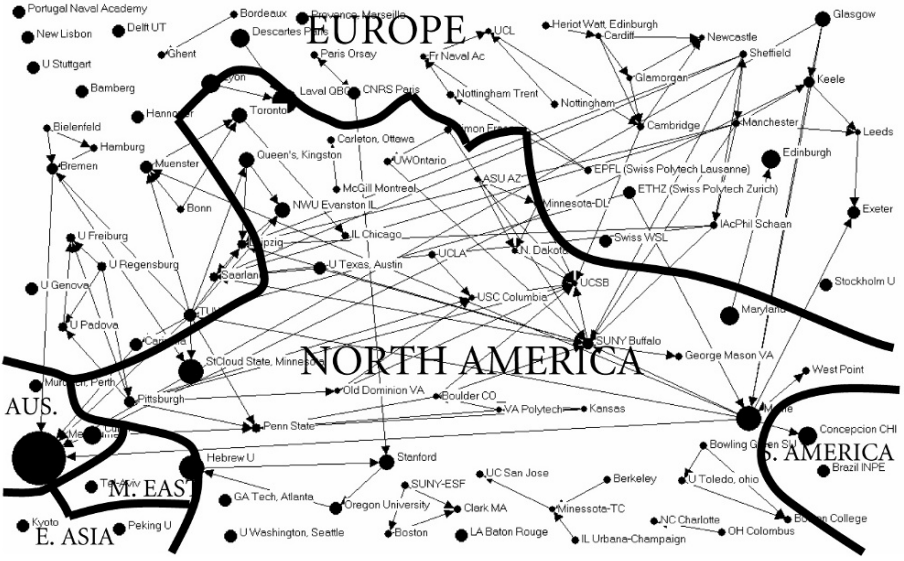


Figure 13.6: Unweighted intracontinental heterogeneity

It is worth noting the changes in patterns of heterogeneity measures for the individual nodes when an intracontinental space is analyzed (Figure 13.6). It is observed that the University of Melbourne maintains its heterogeneity and attractivity in an intracontinental dispersion space, while there is a noticeable difference in TUW, which in this case appears to lose its prominence in terms of heterogeneity. This is because TUW has maximum connections within Europe, having hosted researchers at several points of time in their research trajectories and networks, as compared to those outside Europe. St Cloud State, Minnesota, appears quite heterogeneous, in this case, due to its connectivity primarily to Bremen and TUW, both outside North America, while not supporting any research collaborations or migrations with other universities within the host continent.

Figure 13.7 shows a comparative analysis to that shown in Figure 13.6 using the weighted measure for heterogeneity [cf. Eq. (3)] and the influence on node centralities using this measure, where the homogeneity or heterogeneity of the class is less important as compared to the relative connectivity of the node itself. The University of Maine appears significant because of the number of connections and networks that it hosts both internationally and within the United States. TUW appears prominent and has higher magnitude than that in Figure 13.6, as in this case the relative geographic locations of the connected nodes are considered less important than the number of connections itself. St Cloud, Minnesota, on the other hand, seems to have "lost" its centrality as the numbers of connections to and from this node are relatively small. No significant changes appear for the University of Melbourne.

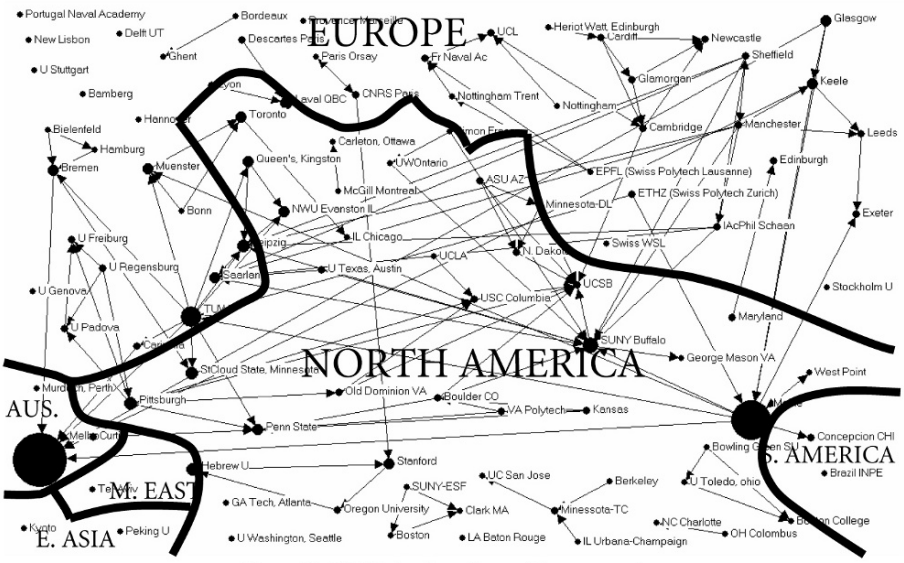


Figure 13.7: Weighted continental heterogeneity

The geographical heterogeneity measures (in both weighed and uweighted versions) support the analysis of the types of connections between the different universities. For instance, it appears that UCSB plays a key role in the United States from a national point of view, as it tends to be connected exclusively to other U.S. universities. This is not the case of SUNY–Buffalo or the University of Maine, as both share, almost exclusively, connections with non-U.S. universities. This highlights the difference in policy (conscious or not) of academic appointments and the university attractiveness abroad. For example, the departments involved in the abovementioned conferences are mainly North American and European, and North America appears more attractive to European academics than the reverse. Nevertheless, and for diverse reasons – of which one can imagine that geographical remoteness plays a role – the East Coast of the United States drags most of the Europeans, who might feel not too far away from home, as they are "just on the other side of the pond". The analysis highlights the way that the geospatial properties of the network are playing a significant role in the development of the nodes' thematic properties and connections in the network.

13.3.3 Similarity Analysis

We complement the graph-based measures by a similarity analysis applied to the two conference series COSIT and GIScience using keywords based on the application of search results from Google, the popular search engine over the Web. Although previous analysis has shown that both conferences have their specific research themes and networks, these are far from being disconnected. The analysis in the previous sections has showed the extent of overlaps between these two research communities and networks. One can then make an attempt at comparing the activity of these networks by measuring their respective footprints left on the Web using the similarity measure proposed in Section 13.2. An immediate measure of conference popularity, and for a given year, is given by the number of Google "hits" de-

rived from a search query defined using appropriate keywords such as the year, name and location of that conference. For instance, the footprint of the COSIT 2005 conference, located in Ellicottville, is evaluated by the search query h(C2005) = h({COSIT + 2005 + Ellicottville}), whereas GIScience 2002 (Boulder, Colorado) is evaluated by the search query h(G2002) = h({GIScience + 2002 + Boulder}). A union query between those two events is then given by h(C2005 ∪ G2002).

Table 13.1: Number of hits returned by search queries

	G2002	G2004	C2001	C2003	C2005
G2002	927	208	35	63	19
G2004	206	740	23	20	22
C2001	38	23	609	58	21
C2003	63	20	58	492	33
C2005	19	22	21	33	393

Table 13.1 shows the number of hits returned on September 19, 2006, whereas Table 13.2 displays the corresponding values returned by the similarity operator. Note the unit value on the main diagonal, since, as one could expect, the self-similarity of a search query is equal to the unit value.

Table 13.2: Similarities between search queries

	G2002	G2004	C2001	C2003	C2005
G2002	1	0.22438	0.037756	0.067961	0.020496
G2004	0.278378	1	0.031081	0.027027	0.02973
C2001	0.062397	0.037767	1	0.095238	0.034483
C2003	0.128049	0.04065	0.117886	1	0.067073
C2005	0.048346	0.05598	0.053435	0.083969	1

These similarity figures reveal several patterns. First, GIScience, although more recent than COSIT, has a much higher visibility on the Web than COSIT. Second, although previous heterogeneity analysis provided evidence of overlaps between these conferences, the similarity analysis shows that the two conferences exist in a relatively independent manner on the Web. Conferences in a series [e.g. Sim(G2002,G2004) = 0.19, Sim(C2001,C2003)= 0.26] are more related than conferences across series [e.g., Sim(C2005, C2003) = 0.09, Sim(C2005,C2001) = 0.07].

Last, more recent conferences have a higher visibility and similarity than the old ones. This is probably a result of the fact that the conference series of COSIT is becoming a de facto standard. These similarity figures are also representative of the quantitative strength of the represented communities and will be useful in future work to analyze the values of the links and overlaps in the networks for individual communities when represented by individual conference series.

13.4 Conclusions and Future Work

In the present digitally connected age, the degrees of separation are becoming lower every day. The World Wide Web provides an important avenue for mapping the underlying semantic and social space that offers many opportunities for generating and studying social networks. The modeling framework proposed in this paper in-

fers and derives a social network from the information on a domain knowledge embedded over the Web. The research communities as supported by the GIScience and COSIT conferences series were used as a demonstrative knowledge domain. Our proposal provides a preliminary contribution towards the inference and exploration of semantic and spatial relationships from the Web-based information space. This approach also adds another dimension to the current research in social networks by combining thematic and geographic operators within such analysis and visualization. Graph, thematic, geographic and similarity operators support a structural and spatial analysis of the network properties. Integration of the spatial and thematic dimensions within the structural analysis permits the correlation of the structural properties with the underlying geographic space. The analysis provided several interesting patterns revealing the universities that play central or outlier roles within these research communities and provided significant indications of the nature of the migratory patterns of researchers as well as collaborative and academic research flows between universities. Most, if not all, of the results in the case study illustrate the theoretical fundamentals of our modeling approach, the intention being not to be exhaustive but representative of the possibility of the framework.

This chapter presents an initial basis for exploring similarity measures between distinct social networks from the Web. This will be further extended to validate the role of the key players within this particular research community and to find the distances and closeness between individual nodes in the research networks. This will be used to develop a more integrated and mathematical profile for the nature of collaborative networks and their respective strengths that form the basis of the network and community involved in geospatial research.

Chapter 14

Participating in the Geospatial Web: Collaborative Mapping, Social Networks and Participatory GIS

L. Jesse Rouse • Susan J. Bergeron • Trevor M. Harris

Abstract. *In 2005, Google, Microsoft and Yahoo! released free Web mapping applications that opened up digital mapping to mainstream Internet users. Importantly, these companies also released free APIs for their platforms, allowing users to geo-locate and map their own data. These initiatives have spurred the growth of the Geospatial Web and represent spatially aware online communities and new ways of enabling communities to share information from the bottom up. This chapter explores how the emerging Geospatial Web can meet some of the fundamental needs of Participatory GIS projects to incorporate local knowledge into GIS, as well as promote public access and collaborative mapping.*

14.1 Introduction

Enabling open access to spatial information has been a research focus in Geographic Information Science (GIScience) since the early 1990s, when a series of debates about the implications of Geographic Information Systems (GIS) technology raised issues of community empowerment, data access, public participation and the incorporation of local knowledge into expert-driven systems (Harris and Weiner 1998). These "GIS and Society" meetings and publications led to initiatives in Public Participation GIS, which has since evolved into the more precisely termed Participatory GIS (PGIS). The basic precept of PGIS is the empowerment of communities through the facilitation of greater community input and access to geospatial data and technologies, community mapping and spatial analysis in support of project decision making.

The availability of free Web mapping applications may now help break down many of the long-standing barriers to the public use of geospatial technologies. Anyone with access to an Internet-enabled computer or mobile device now has the ability to display and interpret geospatial data and even add to that information without expert intervention. Some of the earliest and most notable examples of this can be seen in the collaborative efforts to map the flooding and levee breaks in New Orleans using Google Maps and Google Earth in the aftermath of Hurricane Katrina (Ewalt 2005). The growing number of Web applications that combine user content with free base geospatial data layers clearly demonstrates that broad public access to the Geospatial Web is already in progress. However, the linkage between the core ideas that framed the development of PGIS, and the technologies and techniques enabled by the Geospatial Web, has yet to be considered. This chapter will briefly review the concepts of PGIS and explore how the emerging Geospatial Web might address these goals by promoting public access to geospatial information and tools in support of collaborative community mapping efforts.

14.2 Participatory GIS

The debate about the impact of GIS on society and vice versa in the early 1990s exposed a number of issues related to the use of GIS, including differential access to geospatial data and technologies, and the necessary GIS expertise (Mark 1993; Crampton 1995; Harris et al. 1995; Pickles 1999). This emerging critical perspective recognized that GIS was more than a technical exercise but was situated within cultural, social, economic and ethical contexts (Chrisman 1987). Ultimately, this critical approach led to a move beyond the term "Geographic Information System", which represents a tool or method, to a more appropriate and deeper consideration represented by the term "Geographic Information Science" (GIScience).

As GIS proponents and researchers addressed the issues raised by the GIS and Society debate, so the intriguing possibility of an alternative GIS was considered (Schroeder 1996). The National Center for Geographic Information and Analysis's Initiative 19 investigated the idea of an alternative "GIS2" that would incorporate important aspects of the social-theoretical critique of GIS into a revised technology (Harris and Weiner 1996). Based on the results of the workshop, GIS2 was partially reformulated as a Public Participation GIS (PPGIS) that would specifically address issues of knowledge distortion, differential access to spatial data and geospatial technologies, and community empowerment. In contrast to the still prevalent top-down nature of expert-driven GIS, this bottom-up approach required that local knowledge in the form of oral histories, mental maps, sketches, sound and other qualitative data types be included in the GIS (Shiffer 1999). Many of the recent developments in geospatial technologies now provide support for a more popular usage of geospatial technology that begins to overcome many of the issues identified during the early GIS and Society debates (Goodchild 2004; Pickles 1997; Sieber 2006).

In one of the early examples of PGIS, Harris and Weiner demonstrated how a community-integrated GIS could be developed to address knowledge distortion and differential access through the incorporation of local knowledge into a multimedia GIS. By asking members of the local community in Keipersol, South Africa, to create mental maps of the areas surrounding their villages and then incorporating those maps, text, oral narratives and photographs into a GIS, the community was able to contribute to the land reform process (Harris et al. 1995; Harris and Weiner 1998). However, substantive issues remained related to the silences in the data, who represented the community, whose voices were heard and the impact of socially differentiated local knowledge. The use of the term "community-integrated GIS" was to acknowledge that while community data may be incorporated within a GIS to address structural knowledge distortion, nonetheless, outside experts and resources were still required to take the lead in the design, development and implementation of the PGIS.

14.3 Web-based PGIS

At the same time that PGIS has been evolving, so too has the World Wide Web undergone a rapid evolution. Researchers quickly realized the potential of Web-based solutions for PGIS and GIS2 (Krygier 1999; Kingston 2002). The increasing availability of desktop computers with Internet connectivity provided the opportunity to broaden access to PGIS projects by removing the physical constraint of hosting the GIS in a central location and enabling community members to interact with the GIS remotely (Kingston 2002). In addition, Internet-based GIS greatly facilitates the

incorporation of multimedia such as photographs, drawings, video and audio into GIS (Harris and Weiner 2002). An example of a Web-based PGIS can be seen in the Virtual Slaithwaite project, which utilizes online participation to allow users to individually explore the PGIS and redress issues of data access and the constraints of speaking in the often-charged atmosphere of public meetings. These experiments demonstrated that, although some participants did not feel comfortable using the technology or were unable to gain Internet access, online PGIS in general increased public participation in the decision-making process (Kingston 2002).

Even though Web-based PGIS can potentially broaden community access to GIS, these systems still require substantial expertise and technology to implement them. Internet-based systems are often built on proprietary Web mapping platforms. The successful Community Mapping Network project utilized commercial GIS software to develop online mapping tools that integrated Web mapping and local knowledge to support collaborative community planning in British Columbia, Canada (Mason and Dragicevic 2006). Open-source Web mapping solutions require Web servers and Internet connectivity to store and serve Web-based PGIS projects. Furthermore, many open-source GIS and mapping applications require expertise in software development in order to implement them (Kishor and Ventura 2006). Consequently, while a number of expert-driven Web-based PGIS projects have been successful in broadening access to GIS information and increasing community collaboration in planning and decision making, they are limited in number due to the resources and expertise required.

14.4 PGIS and the Geospatial Web

To date, academic research into the Geospatial Web has largely focused on the technical aspects of spatially enabling the Internet, in order to allow information to be searched for and retrieved on the Web using location as a parameter (Egenhofer 2002). One of the main issues in developing a Geospatial Web has been in building common standards and protocols that allow location to be a linking factor in querying and presenting Web search results. The formation of the Open Geospatial Consortium, Inc., for example, helped to systematize and spur efforts to improve interoperability, as has use of GML (Geography Markup Language) and the open standards for Web Map Services and Web Feature Services.[61] Until 2005, most of the development towards a Geospatial Web was driven by government, academic and industry applications, while the public remained largely unaware of the potential of Web-based mapping and geospatial applications. However, Google Maps and other free Web mapping applications have generated considerable interest beyond the technical creation of a Geospatial Web, and users have focused on how this enabling technology might support a variety of projects, including collaborative mapping and PGIS (Kishor and Ventura 2006).

14.4.1 Web 2.0 and Mapping Mash-Ups

The ability to incorporate user-generated data has transformed the Web 2.0 Internet model into a networked platform capable of playing a key role in the Geospatial Web (Erle et al. 2005; Erle and Gibson 2006). Developers have utilized innovations in Web programming to create easily customizable mapping platforms that provide base data layers such as road networks and aerial imagery. Using these base layers, users can generate custom map applications that combine their own data with the base cartographic data. Known as *mash-ups*, these applications are increasingly con-

tributing to the social networks that are a defining element of Web 2.0. While the social aspects of these applications are generated in different ways, it is the ability to create tags for information with keywords and locations that perhaps has had the most significant impact on these applications.

One of the immediate impacts of the public availability of free Web mapping platforms has been the upsurge in interest in geospatial technologies outside more traditional academic and industry circles. The mash-up phenomenon has demonstrated that, given access to the tools, users from a wide range of backgrounds will create Web mapping applications that link location to a variety of data sets ranging from places that individuals visited, to real-time public transit locations[62] and crime incident locations.[63] Web mapping sites such as MapQuest, Map24, Google Maps and Yahoo! Maps provide access to a wealth of spatial information for much of the world. The associated Web mapping APIs can be used freely by anyone with the technical knowledge to create a Web site and customized maps through a mash-up of their data with cartographic and imagery layers provided by Web mapping sites. As has been recently suggested on CNET:

> *Online mapping is evolving into a historic nexus of disparate technologies and communities that is changing the fundamental use of the Internet, as well as redefining the concept of maps in our culture. Along the way, mapping mash-ups are providing perhaps the clearest idea yet of commercial applications for the generation of so-called social technologies they represent (Elinor Mills, CNET, Nov. 17, 2005).*

While some issues remain regarding the level of knowledge needed to create and interact with geospatial data through these new tools, Web mapping APIs offer a significant step towards an open and accessible geospatial framework. PGIS researchers have long assumed that "the evolving generation of Internet mapping systems will probably play a significant role in future PPGIS projects" (Weiner et al. 2002, 12). Many researchers have created projects that use the Internet for the dissemination of information such as parcel data (Krygier 1999), or capturing comments about communities (Kingston 2005), or even recording public comment within Environmental Impact Asssessments (Harris et al. 2002). Significantly, with the surge in the Geospatial Web over the last year, GIS projects have moved out of the classroom and into broader circulation.

14.4.2 Collaborative Mapping

Perhaps the most obvious aspect of PGIS that is coming to fruition via the Geospatial Web is that of community empowerment through collaborative mapping. Individuals and groups are now able to gather and add data to a central Web mapping platform. By working together, users are able to create and collate far more data than any single individual or group could generate, at little or no cost to the community. The power of such collaboration can be seen in projects like OpenStreetMap,[64] a U.K.-based collaborative project started in 2004 whose goal is primarily to create a global data set of street maps that are not constrained by proprietary or copyright restrictions and could be accessed, edited and customized by users. In addition, the free base cartographic and imagery layers provided by Google Maps, Yahoo! Maps and Microsoft's Virtual Earth have also opened significant new possibilities for collaborative mapping via the Internet, especially since, unlike the United States, many countries do not provide free geospatial information to the public.

Even in the United States, the types of geospatial data available have been largely determined by the needs of the agencies that collect and provide the available data.

These new Web mapping tools have also made tremendous strides in raising awareness among the general public about geography, geospatial information and geovisualization. A number of projects have specifically focused on places that have meaning for different people, and these powerful multiple perspectives can inform us all. Platial[65] is a collaborative mapping and social networking Web site that uses Google Maps to enable users to add content about locations around the world and provide their own "spatial stories" (Aitken 2002). The popularity of Platial and other social mapping sites has led to the coining of a new term, "neogeography", to describe user-generated content that is added to a central Web mapping platform (Turner 2006; Jackson 2006; Newitz 2006).

14.4.3 Virtual Globes and PGIS

It is also important to look beyond the 2D concepts that have driven much of PGIS to date and consider the potential impact of virtual globe applications such as Google Earth, NASA's WorldWind, Microsoft's Virtual Earth 3D and similar technologies. David Maguire (2006) has referred to these virtual globes as Geographic Exploration Systems (GES), which seems appropriate given the capabilities of these systems to view the world in 3D, quickly move to any location in the world, shift between scales and add and remove information. These virtual globes offer an innovative way to connect communities, include local information and allow the material to be explored by many users.

Virtual globes, like their 2D Web mapping counterparts, also provide free base data such as aerial and satellite imagery and elevation layers to provide a contemporary map backdrop to user-generated data. Previously, these data were only available to users of GIS systems. The provision of national and international imagery is an important feature for community-based mapping projects. The Tracks4Africa project, for example, is a community initiative to re-map rural Africa using Google Earth, GPS and documentation provided by community members.[66] Perhaps the best recent example of using virtual globes and the Geospatial Web for PGIS is the collaboration between Amazon Indians and the nonprofit Amazon Conservation Team.[67] Using a variety of geospatial technologies, including GPS and Google Earth, native Amazon peoples are actively participating in the mapping of tropical rainforests in their homeland in order to monitor activities such as mining and clear-cutting that impact their local environment.

14.5 Discussion

As the Geospatial Web becomes more ubiquitous, it is time to examine the prospects for a spatially enabled Web capable of supporting sustainable PGIS projects. This chapter has given examples that demonstrate how the rapidly growing Geospatial Web is providing important services for PGIS. There still remains, however, the need to consider how Web mapping and social networking might also support broad PGIS needs. Until now, PGIS projects have focused on physical communities associated by locational proximity and defined by some areal unit. The Geospatial Web opens the possibility of virtual communities being formed that are united more by a cause or common philosophy than by geography.

The importance of local knowledge, data access, the representation of multiple realities and community mapping are all important concepts in PGIS (Weiner et al.

2002), and it is in these areas that the Geospatial Web is making the most noticeable contributions. Projects like Platial and OpenStreetMap have demonstrated that nonexpert users are willing to initiate and participate in collaborative mapping to share personal information. As the number of such applications grows, so more Internet users will become comfortable with collaborative mapping and the generation of evolving and self-validating data sets.

However, there remain a number of issues in the development of the Geospatial Web and its application to PGIS and community mapping. One of the most significant is the issue of how the community itself is defined. Even in place-based PGIS, it has proven difficult to define the communities and stakeholders involved in the decision-making process. The Geospatial Web introduces another level of complexity, in that the Internet removes the limitations of physical space and allows for the formation of "placeless" communities. Furthermore, although access to the Internet is expanding worldwide, there still remains a significant digital divide that prevents already marginalized communities from participating in the Geospatial Web. Future efforts will need to take these disparities into account and work towards greater access to Internet technologies. The rapid growth in the wireless Internet may facilitate this process in that countries are able to leapfrog the need for an expensive wired infrastructure by tapping into developments in wide-area wireless networking. As well as increasing interest in location-based technologies, the Geospatial Web is contributing to an exponential growth in the amount of spatial information accessible through the Internet. This explosive growth has spurred efforts to monetize Web 2.0 applications and the data that drive them. As a result, community empowerment and participation in the Geospatial Web may, in the end, be limited again by the availability of affordable data.

The Geospatial Web offers significant potential as a platform to support PGIS projects throughout the world. This support comes in the form of Web-based mapping applications, available spatial data, social networking and geographic exploration systems. These collaborative, interactive and easy-to-use environments are attractive platforms for PGIS users who seek to create Internet-based systems that can reach a broad community of users. In the end, with the spread of spatial understanding through new Geospatial Web applications, participation in decision making and the ability to gain information through community feedback will likely see continued growth.

Chapter 15

Sharing, Discovering and Browsing Geotagged Pictures on the World Wide Web

Carlo Torniai • Steve Battle • Steve Cayzer

Abstract. *In recent years the availability of GPS devices and the development in Web technologies have produced a considerable growth in geographical applications available on the Web. In particular, the growing popularity of digital photography and photo sharing services has opened the way to a myriad of possible applications related to geotagged pictures. In this work we present an overview of the creation, sharing and use of geotagged pictures. We propose an approach to providing a new browsing experience of photo collections based on location and heading information metadata.*

15.1 Introduction

With the growing popularity of digital photography, there is now a vast resource of publicly available photos. The availability of cheap GPS devices has made it easy to classify, organize and share *geotagged* pictures on the Web. Geotagging (or geocoding) is the process of adding geographical identification metadata to resources (Web sites, RSS feed, images or videos). The metadata usually consist of latitude and longitude coordinates, but they may also include altitude, camera heading direction and place names.

There has recently been a dramatic increase in the number of people using geo-location information for tagging pictures. The result of a query for pictures with *geo:lat* tag uploaded in Flickr[31] returns 16,048 results between October 2003 and October 2004, 89,514 results for the following year and 171,574 results for the period from October 2005 to October 2006. In principle, the availability of geotagged pictures allows a user to access photos relevant to his or her current location. However, in practice, there is a dearth of methods for discovering and linking such spatially (and perhaps socially) related photographs. In this chapter we focus on geotagged pictures, describing how to add geo-location information to pictures, how geotagged pictures can be organized and shared on the Web and what kind of applications can be built using pictures provided with geo-location information.

The chapter is organized as follows: services and applications related to geotagged pictures are described in Section 15.2; our approach to using geotagged pictures is presented in Section 15.3; possible metadata and distributed environment enhancements, together with benefits and drawbacks of the proposed approach, are discussed in Section 15.6, while Section 15.7 provides conclusions and future work.

15.2 How to Create, Share and Use Geotagged Pictures

In the Web community, geotagging is becomingly increasingly prevalent in photo sharing services that allow users to add metadata, including geo-location information, to pictures. The generated metadata are then used to classify and retrieve im-

ages. Once pictures are geotagged, different kinds of applications can be developed in order to present relations among them and explore new ways of browsing pictures. In this section we discuss services providing tools for geotagging pictures, and applications that use geotagged resources.

15.2.1 Tools for Geotagging Pictures

Flickr is perhaps the premier photo sharing Web site at the time of writing. Following the increasing number of pictures that are manually geotagged by users, Flickr has recently launched its own service for adding latitude and longitude information to a picture. The tool allows a user to select on a map the location in which a picture is taken, and then the corresponding latitude and longitude information is added as metadata to the picture. The process of manual geotagging is quite lengthy, especially the first time we look for a location. The service uses Yahoo! Maps, and the accuracy in the location specification is not fine enough to identify the precise point in which a picture has been taken. In addition, the process doesn't add the latitude and longitude to the picture as standard *geo:long* and *geo:lat* tags nor as EXIF information but rather in an unknown format decoupled from the picture. On the other hand, pictures already geotagged manually with the proper *geo:long* and *geo:lat* values can be automatically referenced on the map.

Zooomr[68] is another photo sharing service that provides a geotagging tool. If a picture with EXIF information on latitude and longitude is uploaded, it is automatically placed on the map. The process of manually geo-referencing an image is similar to Flickr, but here Google Maps is used, providing a more accurate and satisfying geotagging process.

Picasa[69] is a desktop application for organizing digital photos. Recently, a beta version of Picasa (Picasa Web Albums) with a geotagging service integrated with Google Earth[2] was released. Google Earth is used to select the location in which the pictures have been taken, and the latitude and longitude information is added to their EXIF metadata. This tool is very user-friendly and effective, taking advantage of the powerful Google Earth desktop application.

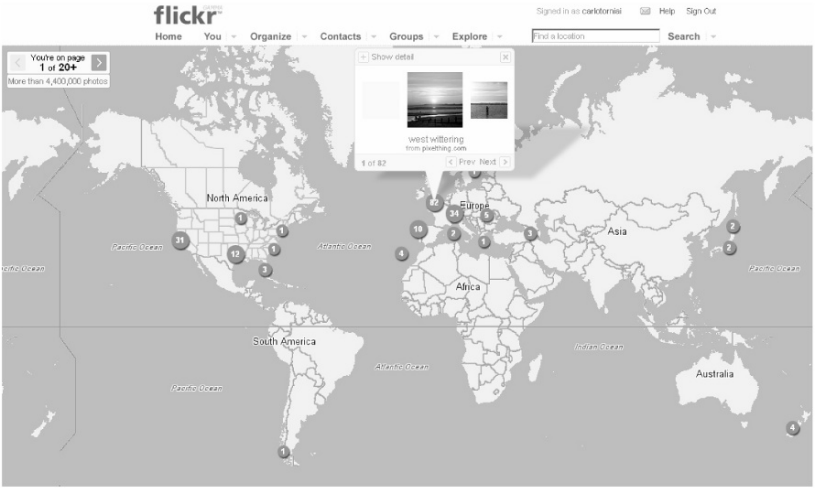


Figure 15.1: Flickr geotagged images browser

15.2.2 Applications Using Geotagged Pictures

The applications for geotagged pictures available on Flickr provide a view of nearby pictures and a browser for geotagged pictures.[70] When looking at a picture on the map, the option "Explore this map" is available and clusters of nearby pictures are displayed. Similarly, in the geotagged images browser a world map with clusters of geotagged pictures is presented. Clicking on a cluster shows thumbnails of the contained pictures (see Figure 15.1).

Zooomr provides a similar application for visualizing pictures on a map. The "browse nearby pictures" feature presents both a map and a textual navigation based on pictures clustered according to the distance from the current picture (see Figure 15.2).

Figure 15.2: Zooomr nearby pictures view

Picasa, as mentioned, uses Google Earth to visualize geotagged images. It can also be combined with Flickr or Zooomr to upload already geotagged pictures. Other Web-based services for geotagging pictures are available. Zoto[71] provides services similar to Flickr and Zooomr but with less features, while jpgEarth[72] allows users to upload pictures related to a location picked up from a Google map, but no search or clustering features are available.

15.2.3 Interaction with Geotagged Pictures

The services and applications described so far provide tools for geotagging pictures and applications that use geotagged data to obtain a cluster view of images on a map, or to find nearby pictures. Other interesting applications take advantage of geotagged resources, building new paradigms of interaction.

Loc.alize.us,[73] a service built on top on Flickr and Google Maps, displays geotagged pictures and provides tool for geotagging pictures and uploading them directly into Flickr. The interesting feature is the possibility to interact with tags and users in order to create and share custom "views" of maps, users and related pic-

tures. Other interaction possibilities are provided by Flickr-based greasemonkey[74] scripts, which enable browsing of pictures based on location information. GeoRadar is a script to search closest photos. A radar screen is displayed in the picture page, and green points on the radar indicate the locations of nearby photos. Thumbnails of nearby pictures are displayed in order of distance from current photo; clicking on a thumbnail causes the corresponding green point on the radar to turn red and a small compass to appear showing the direction from the current picture to the one selected (see Figure 15.3a). Flickr Photo Compass is another script that displays the eight closest photos to the actual one in the cardinal and intercardinal directions: N, NE, E, SE, S, SW, W, NW. By clicking on the direction icons, the user can move around and find other photos (see Figure 15.3b).

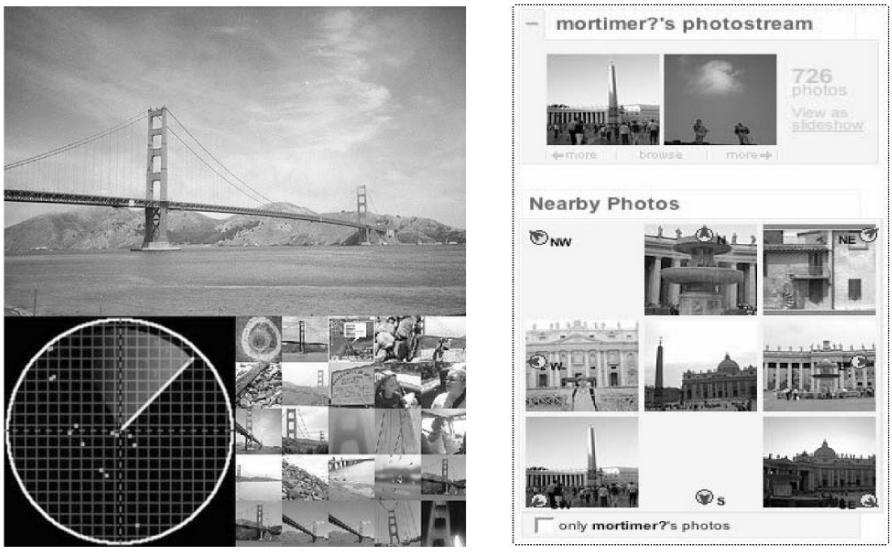


Figure 15.3: (a) GeoRadar screenshot; (b) Photo Compass screenshot

Table 15.1 presents an overview of these services and applications. Notice that most of the services and applications are related to one community and to one service (Flickr). The main mode of interaction is to locate pictures on, and browsed using, a map. In our view this is only one of the potential benefits of geo-referenced data, and we discuss in the next section some recent research projects that use geo-location information to create novel photo browsing experiences.

15.2.4 Research on Geotagged Pictures

In Sharing places,[75] multimedia annotation (photo, video and audio) is associated with physical locations to create a *mediascape*. These trails, based on GPS information and enriched with annotations, can be accessed over the Web or downloaded to a suitable device (e.g., PDA) and experienced in the real world. The trails can be tagged, published for others to find, remixed and shared.

Images are arranged according to their location in the World-Wide Media Exchange (Toyama et al. 2003) while time and location are used to cluster images in PhotoCompas (Naaman et al. 2003). Realityflythrough (McCurdy and Grishwold 2005) presents a very friendly user interface for browsing video from camcorders

equipped with GPS and tilt sensors, and a method for retrieving images using proximity to a virtual camera is presented in Kadobayashi and Tanaka (2005).

In Photo Tourism (Snavely et al. 2006) a system for interactively browsing and exploring large unstructured collections of photographs is presented. Using a computer vision-based modeling system, photographers' location and orientation are computed along with a sparse 3D geometric representation of the scene. Full 3D navigation and exploration of the set of images and world geometry, along with auxiliary information such as overhead maps and geo locations, are provided by the photo explorer interface.

These approaches provide a user experience enhanced by geo-information but don't rely on standard format for metadata nor provide a distributed environment for exchanging metadata. As already pointed out (Cayzer and Butler 2004), we believe that metadata related to pictures and their locations should be expressed in a common and sharable standard so that they may be used by other applications. Sharing picture metadata across a distributed environment using an open standard such as RDF (W3C-RDF 2002) can lead to interesting evolutions in the way in which pictures and other multimedia geotagged content are shared, discovered and browsed.

Table 15.1: Geotagged images applications and services overview

Applications/ Services	Goal	Geo-Related Services	Standard Format	Other Services and Technologies
Flickr	Photo sharing	Geotagging tool, geotagged picture browser	None	Yahoo! Maps
Zooomr	Photo sharing	Geotagging tool, geotagged picture browser	None	Google Maps
Picasa	Photo organizer	Geotagging tool	EXIF	Google Earth
Loc.alize.us	Geotagged pictures browsing service	Geotagging tool Social network related to picture	None	Flickr Google Maps
GeoRadar	Enhanced Flickr interaction	Location-based image browsing	None	Flickr Greasemonkey
Photo Compass	Enhanced Flickr interaction	Location-based image browsing	None	Flickr Greasemonkey

15.3 Building Applications with Geotagged Pictures

Our contribution in applications related to geotagged pictures explores the kinds of metadata that can be captured at the time a photo is taken and ways to link photos together according to this metadata. The objective of our work is to create an experience where someone can view a photo on the Web and then jump to other photos in the field of view or taken nearby. It draws on the network effect of the Web by including not only the user's own photos but any photo that can be discovered with suitable metadata. This includes location (GPS or other mobile location) and heading information to identify the position and direction of the camera. The photos discovered may have been taken by different people and are shared on the Web. The key to this linking is location and heading metadata attached to the photo. There are no explicit hyperlinks between photos, making it easy for people to contribute. Automatic linking is achieved by the discovery of photos on the *Semantic Web*.

The main idea is to capture RDF metadata related to pictures and photo collections and share these descriptions in a distributed environment. Spatial relations between nearby pictures are discovered by means of inference over their RDF descriptions. We have implemented a proof of concept system comprising the algorithm for inferring spatial relations between different pictures (see Section 15.4), a distributed system for sharing metadata and picture discovery and a Web client that uses these RDF descriptions to provide a browsable interface, allowing users to explore shared photo collections through their spatial relationships with each other (see Section 15.5). To define the structure and the content of metadata for picture description, we consider the existing RDF schemata that capture the following information: latitude, longitude, heading information, author, date and time, title, annotation about location and EXIF metadata.

We used both an RDF translation of the EXIF standard (W3C-Exif 2003) and Basic Geo vocabulary (W3C 2003) for latitude and longitude. Heading information and camera-related data (focal length, focal plane resolution and so on) are expressed using the RDF format of the EXIF standard. Dublin Core[18] was selected for defining author, title, date, time and annotation about location.

To describe the location context, we used the Dublin Core *dc:coverage* tag. The purpose of *dc:coverage* is to define the extent or scope of the content of a resource and typically includes spatial location (a place name or geographic coordinates), temporal period (a period label, date or date range) or jurisdiction (such as a named administrative entity). Additionally, we introduced a hierarchical order into the values of this tag, namely place or area, city and country. For instance, values representing a picture taken at the Watershed in Bristol would be, "Watershed, Bristol, UK". Furthermore, this hierarchical tag could be used to generate a less specific tag, "Bristol, UK", providing more flexibility in the discovery process.

A collection of pictures is expressed in RDF as a list of images with a title and a creator expressed through the *dc:creator* and the *dc:title* tags.

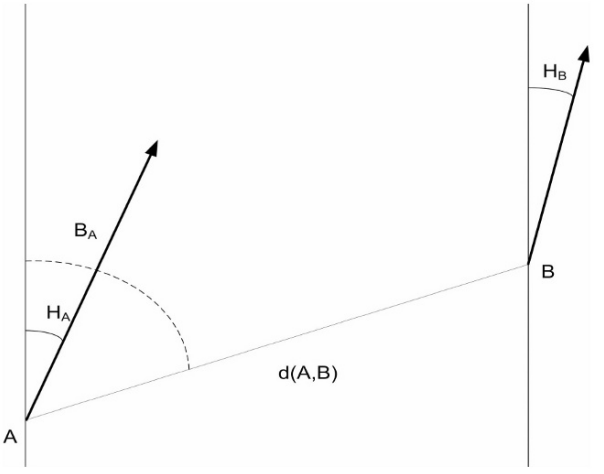


Figure 15.4: Field of view evaluation; if $|H_A - B_A|$ is less than a given threshold, point B is in the field of view of point A; if $|H_A - H_B|$ is less than a given threshold, then the pictures have a similar heading; if these conditions are met, then $image_B$, taken at B, is in field of view of $image_A$ taken at A

15.4 Discovering Pictures' Relations

RDF descriptions capture the spatial relationships between pictures. We define a simple algorithm that extracts the following information: (i) field-of-view evaluation (moving forward – zoom); (ii) spatial relations (turning – pan).

The field-of-view relation describes the fact that from a picture taken at A ($image_A$) one can *move towards* the picture taken at B ($image_B$). The way in which the field of view is evaluated is shown in Figure 15.4. This states that for $image_B$ to be in the field of view of $image_A$, one must be able to see point B in $image_A$, and $image_B$ must have a similar heading direction to $image_A$.

The method for field-of-view evaluation is shown in Algorithm 1. FOV_THRESHOLD has been set to 150 m, while the bearing angle threshold T_{bear} and the heading direction threshold T_{head} have been heuristically set to 20 degrees.

Algorithm 1. Field-of-view evaluation algorithm

```
for each image pair (image_A, image_B) in the collection
  evaluate distance d(A, B)                    // distance between A and B
  if d(A, B) < FOV_THRESHOLD then
      evaluate B_A                             // bearing angle between A and B
      if (|H_A - B_A|< T_bear )                 // i.e., point B can be seen in im-
age_A
         AND (|H_A - H_B| < T_head) then        // i.e., image_A/B have similar head-
        ings
         set fov_relation(image_A, image_B)
```

Spatial relations refer to the direction in which you have to turn, standing in A, in order to see the picture taken at B. If the pictures $image_A$ and $image_B$ have been taken within a given range of each other, we consider the pictures to be taken at the same location so that their relative spatial position is given by the difference between their heading information. Referring to Figure 15.5, we can say that you can turn right from A to B.

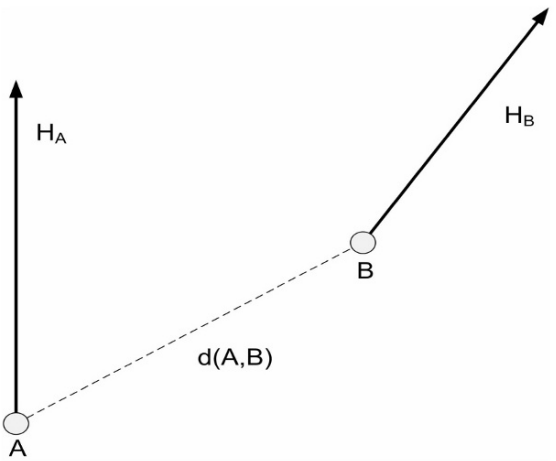


Figure 15.5: Spatial relation evaluation; if d(A, B) is less than a given threshold, then the spatial relation is given by ($H_A - H_B$)

The algorithm for spatial relation discovering is shown in Algorithm 2. DISTANCE_THRESHOLD has been set to 15 m, taking into account the GPS accuracy.

Algorithm 2. Spatial relations discovering algorithm

```
for each image pair (image_A, image_B) in the collection
 evaluate distance d(A, B)                    // distance between A and B
 if d(A, B) < DISTANCE_THRESHOLD then
  diff_angle = H_A - H_B
  case diff_angle
        0 to +22.5 OR -337.6 to -360          : position = Front
        +22.6 to +67.5 OR -292.6 to -337.5    : position = Front_Right
        +67.6 to +112.5 OR -247.6 to -292.5   : position = Right
        +112.6 to +157.5 OR -202.6 to -247.5  : position = Back_Right
        +157.6 to +202.5 OR -157.6 to -202.5  : position = Back
        +202.6 to +247.5 OR -112.6 to -157.5  : position = Back_Left
        +247.6 to +292.5 OR -67.6 to -112.5   : position = Left
        +292.6 to +337.5 OR -22.6 to -67.5    : position = Front_Left
        +337.6 to +360 OR -0.1 to -22.5       : position = Front
 set spatial_relation(position, image_A, image_B)
```

The output of the algorithm is an RDF model describing the relations discovered between the pictures. We have defined simple properties describing the field of view (*has_in_fov*) and spatial relations (*Front*, *Left*, *Right*, *Back_Left*, *Front_Right* and so on).

15.5 Distributed Environment

A distributed test environment has been implemented in order to evaluate the pictures discovering process and the algorithm for relations evaluation across different photo collections. This environment is composed of a set of "clients". Each client exposes its photo collection(s) (i.e., the RDF collection descriptions files) to its peers by means of SPARQL (W3C 2006a) endpoint(s). The clients hold, but do not need to share, the inferred spatial relations between pictures.

The process of discovering related pictures is described in Algorithm 3. Discovery is performed through queries against remote clients and does not require the relatively expensive computation of spatial relations. Instead, photos are selected by their coverage, expressed as relatively simple location hierarchies.

Algorithm 3. Pictures discovering algorithm

```
expand the coverage tags in the collection
for each distinct coverage
  for each client
      query client for coverage entries
      evaluate relations(client_collection, virtual_collection)
```

The first step is the expansion of hierarchical *dc:coverage* tags in a client's own collection. This allows an SPARQL query to retrieve photos at varying degrees of granularity. For example, given a picture with the coverage "Peto Bridge, City Center, Bristol, UK", the expanded coverage tags will be the following:

<dc:coverage> Peto Bridge, City Center, Bristol, UK </dc:coverage>
<dc:coverage> City Center, Bristol, UK </dc:coverage>
<dc:coverage>Bristol, UK </dc:coverage>

The client asks other known clients for pictures that have the same coverage entries as the ones related to its own collection. This is performed by means of SPARQL queries against (similarly expanded) *dc:coverage* tags. It would also be possible to use GPS latitude and longitude information in the SPARQL queries, but this would be relatively expensive. As a result of this query process, a list of images is returned to the client. Only when potentially relevant photos have been discovered and their metadata retrieved from a remote client do we begin to evaluate the specific spatial relationships between them. These images can be considered as a virtual collection of images; candidates that may have some relation with the pictures in the client's own photo collection. The client executes the algorithm for relations evaluation between its collection images and the candidate images. Every relationship discovered is added to the RDF model. At the end of this process the client will hold all the relations between its own pictures and pictures of the remote clients.

The distributed environment and the algorithm for relations evaluation permit the growth of the RDF relations model. This holds the information required for building the browser interface for picture collections. The interface is shown in Figure 15.6.

Figure 15.6: Browsing interface

The pictures described in RDF can be accessed by a thumbnail menu or a Google Maps panel. Moving the mouse over the markers on the map causes the latitude, longitude, heading and coverage information for the corresponding picture to be displayed. The user can browse the pictures by means of the navigation arrows surrounding the pictures that show the direction in which a user can move from the perspective of the current picture. Pictures in the field of view can be reached by clicking on the current picture.

For our experiments we used a set of 100 pictures related to 3 different cities. Latitude, longitude and heading information were collected on a Suunto G9[76] watch at the time the pictures were taken and then later injected in the EXIF data for each picture. The RDF collection files were created by a batch program reading the EXIF information directly from the pictures. The test environment was composed of four clients. Each client was implemented using a Joseki[77] SPARQL server running as a Web application under Apache Tomcat. The browsing interface was developed as a Web application using Jena[78] and Velocity.[79]

15.6 Discussion: Alternative Representations, Additional Metadata, Scalable Architecture

In our approach we used the semantic Web recommendation Resource Description Framework (RDF) to describe photo collections and metadata related to the pictures they contain.

Among other metadata formats (EXIF or XML, for instance) RDF was chosen because we want to deal with metadata decoupled from the actual resources in order to be able to store, process and expose the information about pictures (among them the location as the URI of the resource) independently of storing the actual photograph. Moreover, we want to be able to define and extend relations between metadata and have the possibility to take advantage of RDF inference capabilities that are not available in XML. In addition, RDF offers the following advantages:

- RDF is expressly designed to provide a standard, extensible format for machine-readable metadata. RDF is an open standard, allowing widespread deployment and consumption. Using RDF means that metadata can be shared and reused more easily.
- RDF is "syntax–neutral"; different RDF vocabularies all share the same syntax. This allows us to easily mix different vocabularies and load any vocabulary into any tool.
- Ontologies for image metadata are already available in RDF format.

The following ontologies are examples of those that can be used in order to define pictures metadata:

- W3C (W3C 2002) suggests three simple schemata – Dublin Core (for title and description), a technical schema (for camera type and lens) and a content schema (oft-used tags like Baby, Architecture and so on).
- Time can be dealt with as a Dublin Core tag or by treating events as first-class entities (W3C-Cal 2002).
- Space can be described using precise geographical descriptors, like latitude and longitude, and for which ontologies are already available (Section 15.3).[80] To represent hierarchical relations such as "England contains London", we could use formal approaches like the space namespace ontology.[81] A more ambitious, though incomplete, schema based on ISA standards has also been proposed.[82] Differing degrees of accuracy can be catered for by taking a layered approach[83] ("within 10 m," "within 100 m," "within 10 km"…). An alternative approach is to consult a controlled vocabulary with concrete place names.

- Device metadata are often provided within a photo in EXIF format, for which the RDF version exists. Other terms such as focal length relevant to cameras are represented in Morten Frederickson's Photography Vocabulary[84] and in Roger Costello's Camera ontology.[85]
- Topic tags can be mapped to Flickr tags, as the URI for a Flickr tag is simply its URL. The RDF property used to connect a photograph to a Flickr tag would, however, need to be a custom property. The tag hierarchy can be represented within RDF using rdfs:subClassOf or skos:broader.[86]

Our ontology reuses some of these existing ontologies for EXIF and Basic Geo (WGS84 lat/long) metadata. Heading information and camera-related data (focal length, focal plane resolution and so on) are expressed using an RDF version of the EXIF standard. Dublin Core describes author, title, date, time and annotation about location. We have introduced our own vocabulary for defining field-of-view and spatial relations as described in Section 15.4.

Our approach for hierarchically structured locations uses the *dc:coverage* property and the values it may contain. This approach is very lightweight compared to relations defined more formally but has the following advantages:

- simple expression of the "place or area, city, country" order,
- tag-like format that users can easily create,
- more accessible than a series of properties values.

The advantages of letting users define their own vocabulary for classifying information has already been demonstrated by the growth of tagging community, while the effectiveness of folksonomies in information classification and retrieval is becoming more and more relevant. One could extend our approach using constraints on tag-like format of property values, or indeed link photographs using controlled vocabularies. Other metadata can be added to the proposed picture description. In particular, it would be interesting to add social metadata related to pictures so that social relations, other than spatial, can be discovered and presented to the users, providing a *social exploration* of shared picture collections.

Our prototype has been a useful proof of concept but is not yet suitable for real deployment. A P2P architecture would provide an optimization of query caching and routing between the different clients at the expense of complexity in the client implementation. However, a centralized server, which would act as the repository of the pictures' metadata and evaluate the spatial relationships between users' pictures with batch processes, allows the development of a simple Web-based service without the need of a client-side application. This is a lighter-weight solution for users who wouldn't have to download and install a full software application.

Compared to other approaches and applications, our system has the benefit of standard metadata descriptions that can easily be shared and reused in many different applications and services. The browser application built on top of these descriptions is an example of what can be done using our approach. RDF provides flexibility in how spatial information is encoded, processed and computed. One can imagine, for example, a browser based on social networks or an algorithm combining latitude, longitude, coverage and geographic thesauri for more accurate spatial labeling. The lightweight approach proposed for computing picture relations, and indeed the choice to rely purely on metadata rather than on information gathered from heavyweight image processing, makes our solution suitable for real-time and Web-based applications.

15.7 Conclusions

In this chapter we have explored ways to create, share and use geotagged pictures available on the Web. As an example of applications using geotagged pictures, we have implemented a prototype system providing ways to

- share geotagged pictures,
- discover pictures through geotag metadata,
- present geotagged pictures and their spatial relationships.

An algorithm for inferring spatial relations between different pictures using location and compass heading information embedded in the RDF description of the pictures has been presented. A testing environment for metadata sharing and picture discovery has been implemented so that users' photo collections are enhanced by relations with other users' pictures. We have shown how, based on geographical metadata expressed in RDF, it is possible to build a service for discovering, linking and browsing geographically related photos in a new way. Our future work will deal with experiments on large test beds in order to obtain meaningful performance evaluation, improve scalability and improve the user interface.

Chapter 16

Supporting Geo-Semantic Web Communities with the DBin Platform: Use Cases and Perspectives

Giovanni Tummarello • Christian Morbidoni • Michele Nucci • Ernesto Marcheggiani

Abstract*. The aim of this chapter is to show how the need for advanced cooperative annotation and information exchange can be addressed using a paradigm called "Interconnected Geo-Semantic Web Communities". The use cases and its associated needs are highlighted, and then the base tool for this work, the DBin Semantic Web information manager, is focused on. DBin enables users to create and experience the Semantic Web by exchanging RDF knowledge in peer-to-peer (P2P) "topic" channels. Once sufficient information has been collected locally, rich and fast browsing of semantically structured knowledge becomes possible, even offline, without generating external traffic or computational load. DBin has a number of modules to support cooperative tagging and annotations of geographical objects. Different communities of users, e.g., concerned with different kinds of geographic objects, can each exploit DBin to cooperate in enriched geo-semantic spaces. Advanced users, e.g., cultural heritage agencies, can join multiple groups at the same time and use collective cross-domain knowledge.*

16.1 Introduction

The aim of this chapter is to present an innovative Geo-Semantic Web scenario based on a paradigm called "Interconnected Semantic Web Communities". The idea is to enable end users to create and experience the Semantic Web by exchanging structured information in P2P "topic" channels. A Rich Semantic Web Client (RSWC) provides the users with the connectivity to such topic channels and rich domain-specific interfaces.

The information exchanged within such P2P groups is structured according to the Resource Description Framework (RDF),[87] standardized by the World Wide Web Consortium (W3C), and by using appropriate ontologies for each specific domain of interest. In the case of Geo-Semantic Web scenarios, where people are interested in annotating geospatial objects, existing Semantic Web geographical ontologies will be used. However, this must not necessarily be the sole case, as Geo-Semantic Web users might also be interested in annotating events, people, geopolitical entities, etc. By participating in such topic channels, users build and update a local rich knowledge base, which supports high-speed local browsing, searching, personalized filtering and processing of information. Such a knowledge base is then relevant for planning and decision supporting system (DDS).

16.2 Geo-Semantic Web Communities

The Como Lake area, in Italy, is particularly rich in historical, artistic, ecological and touristic landmarks and attracts approximately 3 million visitors per year. Even if the area is welcoming and well organized, the importance of providing tourists with rich and personalized choices is quickly increasing. This requires integration and sharing of information between local actors using state-of-the-art information technologies. In particular, the case study focuses on Villa Mylius Vigoni, a historical villa of international relevance, comprising a large historical park and a romantic garden. Its maintenance involves institutions and organizations ranging from European to local institutional level, each of which performs a number of specific tasks and would benefit from cooperative exchange of geospatial semantic annotations.

The Geo-Semantic Web Community tool, proposed here, is a specific application of the DBin Semantic Web Platform (Tummarello et al. 2006), which enables users communities to annotate concepts of common interests. For this study, DBin has been fitted with both geo-enabled visualization and editing modules as well as with processing capabilities to handle data structured according to geotagging ontologies (such as the W3C Geo Ontology).[53] In this work the term "ontology" refers to a set of classes, relations and constraints defined using Ontology Web Language (OWL)[88] for modeling a specific domain.

Geo-Semantic Web communities are characterized by the typology of knowledge that is exchanged within them. This means that when connecting to a specific community, users will only share and exchange that information relevant to the community itself. In DBin this is achieved by means of topic-specific P2P groups based on the RDFGrowth algorithm discussed in Section 16.3.

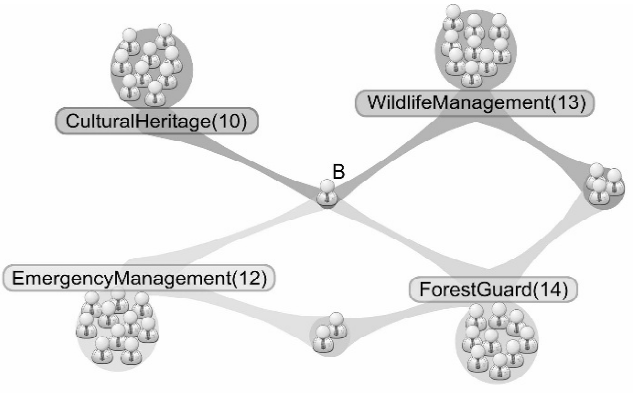


Figure 16.1: Our reference use case: groups of experts make use of the annotations in their own domain; certain experts, by joining multiple communities, can make use of the joint knowledge

In the proposed case study (illustrated in Figure 16.1), cultural heritage agencies are interested in annotating the state of cultural patrimony. At the same time parks and woods maintenance is performed by Forest Guards through specific actions, such as dangerous trees pruning or cutting, service buildings and routes maintenance, etc.

In case of emergency, users like Bob (marked as B in the picture), might join all the groups and collect all the information from all the local Geo-Semantic Web Communities, being able, for eample, to elaborate global maps of risk or special annotations, which he can then share with the others in his specific group.

The proposed RSWC-based solution offers to each group a specific environment where such annotations can be edited, browsed and queried to assist decisions and management. Collaborative metadata annotations created within each community can also point at rich media, posted on the Web (e.g., pictures, documents, long texts, etc.). Users who receive an annotation could then reply or further annotate it, locally, for personal use or for public knowledge.

A novel feature of Semantic Web Communities is that they are naturally interconnected: if two communities share Uniform Resource Identifiers (URIs), an object's annotations, originally posted in a specific community, are automatically cross posted to other communities, as soon as the annotated object is of interest to both the communities.

Enabling users to interact with the RDF data in a natural way is a central and strongly domain-dependent issue, as each community dials with different kinds of information and has different needs in terms of browsing and editing capabilities. Our solution is to provide a way for a "group leader" to define, and make available to the public, a set of domain-specific "interaction profiles" called Brainlets. New Geo-Semantic Web Communities can be started at will, with relatively little technological effort, by creating and publishing both a P2P group and a Brainlet.

16.3 The RDFGrowth P2P Engine: High-Level Overview

DBin clients base on the RDFGrowth algorithm as the main channel for collecting and distributing RDF data. The algorithm is presented and discussed in Tummarello et al. (2004); a high-level overview is given here.

Previous P2P Semantic Web applications, such as (Nejdl et al. 2002, 2003), (Chirita et al. 2004) and (Cai and Frank 2004), have explored interactions among groups of trusted and committed peers. In such systems peers rely on each other to forward query requests and collect and return results. On the contrary, the real-world scenario of peers, where cooperation is relatively frail, is considered here. In RDFGrowth, peers are certainly expected to provide some external service, but commitment is minimal and in a "best effort" fashion: no commitment in terms of complex or time-consuming operations, such as query routing, collecting and merging, is required from peers.

RDFGrowth enable users to aggregate around topics of interest, creating P2P groups. When a user joins a specific group, RDFGrowth begins to collect and share only information of interest within the group. A topic's definition is given by the Group URI Exposing Definition (GUED), an operator that, once downloaded and applied to an RDF data set, returns all and only the resources (URIs) that adhere to a set of semantic constraints (e.g., "trees within 1 km from the centroid of Villa Vigoni"). Peers in the same P2P group exchange information patches about the resources of interest (selected at each peer by the group's GUED), so that, after a transitory period, each of them will know exactly the same about the community domain and will have stored this knowledge locally.

It is to be noticed, however, that the RDFGrowth approach is particularly suited for the scenario of geo–annotations, as it enables the peers to work offline, therefore possibly on mobile devices far from connectivity points.

P2P topic groups can be configured by editing an XML configuration, defining the GUED to be used. A GUED can be implemented as a set of Semantic Web queries, which select resources of interest. In the case of a group interested in trees located in the area of Villa Vigoni, a possible GUED configuration might be the one given in Example 16.1.

```
<group name="Botany">
  <gued query="SELECT X FROM {X} rdf:type {<http://dbin.org/trees#Tree>};
               geo:lat {lat}; geo:long {lng}
               WHERE  lat > 55.893411  AND  lat < 55.988012
               AND    lng > -3.316841  AND  lng < -3.064100" />
  <default_brainlet name="BotanyBrainlet" uri=http://dbin/brailets/botany />

</group>
```

Example 16.1: The botany P2P group configuration

In such a system, which deals with potentially large and unregulated communities, it is important to have authorship information about the annotations received from the network. To enable this, the methodology described in Tummarello et al. (2005) is used, allowing each piece of information inserted by users to be digitally signed and the relative authorship information to be exchanged within a community along with the data itself.

Once authorship of each annotation can be derived, a variety of local filtering rules can be applied at will. For example, users can build a local "black list" policy, in order to hide annotations from untrusted authors.

16.4 Defining Semantic Web Communities: Brainlets

Brainlets can be though of as "configuration packages" preparing DBin to operate on a specific domain. Brainlets are perceived by users as full "domain applications run inside DBin". In short, Brainlets define settings for *ontologies* to be used for annotations in the domain; general layout, defining *UI components* and *interactions among them*; templates for *domain-specific "annotations"*; templates for readily available *"precooked" domain queries*; templates for *wizards for creating new domain objects* (to avoid duplicated URIs, etc.); suggested *trust models* and *information filtering rules*, e.g., identities of "founding members" or authorities.

Technically, Brainlets are Eclipse RCP plug-ins that can be installed into DBin. Creating a Brainlet does not require programming skills, as all the involved tasks are related to knowledge engineering (e.g., selecting the appropriate Ontologies) and editing of an XML configuration file. This section discusses the main points of such a configuration, explicitly referring to the *Botany Brainlet* (Figure 16.2), which, used in conjunction with the P2P group defined in Example 1, enables one of the communities considered in the case study (Section 16.2).

In creating a new Brainlet, an important step is the choice of appropriate ontologies for the domain of interest. Once they have been identified, the corresponding OWL files are usually included and shipped in the Brainlet itself, although they could be placed on the Web. In the case study, existing ontologies have been used to foster information reuse and interoperability.

Although RDF has a graph model, graph-based visualizers have well-known issues in terms of usability. As users are familiar with folder-like structures, authors decided to provide resources browsing based on flexible and dynamic tree structures. Such an approach can be seen to scale very well with respect to the number of resources. The result is a configurable semantic folder tree, which can have multiple branches, each one organizing the same data with respect to different criteria. The Botany Brainlet navigator is shown in the left part of Figure 16.2: the Green Areas branch organizes objects according to their location, while the Tree branch classifies resources by genre and species.

Within a specific domain, there are often some queries that are frequently used to fulfill relevant use cases. For example, in the Botany Brainlet, such a query could be "find all trees of kind X near place Y". The precooked queries facility allows Brainlet creators to provide such "fill in the blanks" queries to end users.

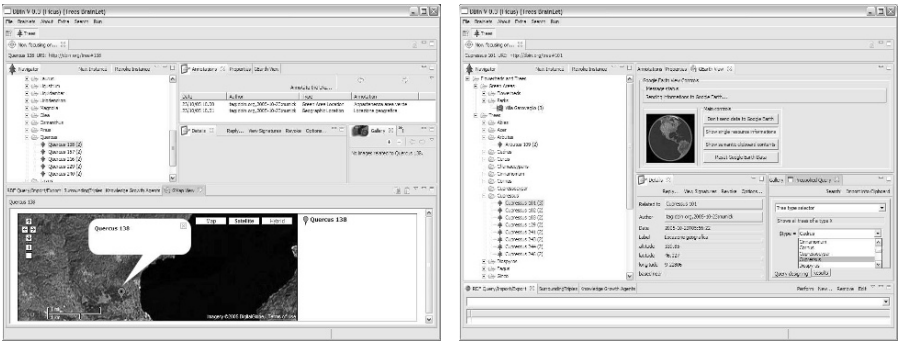


Figure 16.2: [left] Google Maps-based module, which can be used to visualize geographical annotated objects as well as to insert geospatial annotations; [right] UI module to perform geospatial precooked queries and specify settings for the Google Earth plug-in

For fostering knowledge interoperability within communities, a methodology is needed to avoid different users choosing different URIs for identifying the same concept. This can be achieved by defining procedures for assisting the user in assigning identifiers to instances. In the case of geographical objects, a natural identifier can be often derived: houses have an address, other objects, such as trees, have defined spatial coordinates, etc. Such procedures are encoded into what are called URI wizards. By XML configuration it is possible to define customized URI Wizards to assign meaningful URIs to each specific class of resources. Solutions for minting identifiers on the Semantic Web are still in their infancy. The URI wizard approach and the DBin plug-in module offer the flexibility to accommodate future methods as they will be made available and/or reach popularity.

Brainlets use ontologies to assist the users in creating simple annotations (e.g., suggesting which properties can be associated to a resource based on its type). A Brainlet creator can, however, choose to create "complex annotation types" using an ad hoc ontology. Upon selecting "Add advanced annotation" in DBin, the system determines which advanced annotations can be applied to the specified resource and provides a wizard. For example, in the Botany Brainlet, a specific Geographic Annotation has been defined that allows a template with geospatial information (e.g., geographic coordinates) to be filled.

Brainlets can also use additional plug-ins to address domain-specific needs as in the Botany Brainlet, where specific plug-ins provide real-time interaction with Google Earth and Google Maps.

Brainlets, by providing an aggregation medium for ontologies, users and data representation structures, are therefore good catalysts of the overall semantic interoperability process. As users gather around popular Brainlets for their topic of choice, the respective suggested ontologies and data representation practice will form an increasingly important reality, acting as an incentive in using compatible data structures and ontologies when creating a new Brainlet.

16.5 Conclusions and Related Work

Many approaches to the visualization of semantically annotated data have been proposed in the literature. RDFGravity exploits graphs as visualization tools of RDF triples.[46] Previous publications discussed integrated environments for both RDF browsing and authoring (Pietriga 2002). Welkin[89] is an RDF visualizer based on elliptical zooming of the connections among resources, while Longwell[90] implements the idea of faceted browsing of RDF data. MIT Haystack (Quan and Karger 2004), a tool for development of Semantic Web applications, also focuses on the interface organization (layout and functionalities). While pros and cons can be argued for each specific approach, it is clear that user interface issues are complex, with no clear single solution. DBin responds to this by the Brainlet UI interface, which enables a topic-specific mash-up of different visualization paradigms.

DBin stands out as an end-user-centered application that provides Semantic Web capabilities and a plug-in-based, open-source, rich client approach. The combination of these two effectively enables efficient extension and integration into domain-specific tools and ontology-based browsing, querying and searching information. The addition of the proper geographical annotation handling components enables *Geo-Semantic Web Communities*, which can be quickly started by domain experts.

Especially when compared to full-featured Web GIS or specific software, the support for geospatial use cases is today still limited to relatively simple objects, in particular to those that have a well-specified location, that is, those that can be approximated with a point. The purpose of this work, however, it is not to match specific state-of-the-art tools. The potential of Semantic Web-based annotation communities seems unprecedented in terms of flexibility and for the ease of integration of information across communities. This turns out to be very important, for example, when providing decision support in landscape resources participatory planning (Marcheggiani et al. 2007).

It is clear how the full potential of Geo-Semantic Web Communities is to be explored and validated. This is, however, the case for the whole Semantic Web initiative and, with respect to this, the contribution of the DBin platform and the applications that have been shown here is to represent, arguably, the most tangible instantiation of such technologies available for actual use today.

DBin is distributed under the Gnu Public License and includes geographic visualization and editing plug-ins. Further documentation, a screen demo and compiled executables can be found at `http://dbin.org`.

Chapter 17

A Geospatial Web Platform for Natural Hazard Exposure Assessment in the Insurance Sector

Julien Iris • Jérôme Chemitte • Aldo Napoli

Abstract. *The work of natural hazard exposure assessment involves various geographic data sets (referential, hazard, assets) and various disciplines intended for insurance professionals (catastrophe modeling, prevention engineering). The emergence of Geospatial Web technology induces the emergence of new sets of online services. Mission Risques Naturels (MRN) is a French actor in the mutualization and diffusion of information on natural hazards knowledge and prevention for the general interest of insurance professionals. The MRN Web-GIS platform has been built to address these requirements. This chapter starts by presenting the role of MRN in the network of natural hazard assessment. It then presents Geospatial Web tools for natural hazard exposure assessment as well as the system architecture of the MRN Web-GIS platform, including all its services.*

17.1 Introduction

Geographic data sets are intended to be used when spatial relationships have an essential relevance. Obviously, this is the case in the natural risks field, as they play a crucial part in providing qualitative and quantitative information for

- local and national authorities to justify the means allocated to the prevention and assessment of economic disorders,
- territorial risk managers to be able to argue about the structural protection measurements (dikes, retaining tanks, etc.),
- crisis managers to identify the critical zones on which to concentrate their efforts,
- individuals to estimate their need for the purchase of an insurance policy, investments in mitigation measurements and collective protection measures,
- reinsurance and insurance companies to evaluate risk accumulations, to incite prevention, to know the exposure of their portfolio or to refine their policy of subscription (according to the national legal framework).

The association "Mission Risques Naturels" (MRN) was created by the French federation of insurance companies (FFSA) and the mutual insurance companies group (GEMA) in 2000 after a year of several natural catastrophes (floods and storms). Its object is contained in its full title: mission of the insurance companies for the knowledge and prevention of natural risks. Indeed, for the insurance profession, it consists of participating in the construction of better knowledge on natural risks and caring about technical contributions to prevention policies (Nussbaum 2000). MRN is an interface with the various stakeholders of natural risk management. Above all its role is to concentrate, organize and restore rough information and diffuse added value contributions.

This chapter presents the interactions between the Geospatial Web concepts and the natural hazard exposure assessment. Section 17.2 presents an introduction to the modeling of natural hazard exposure assessment using geospatial tools. Section 17.3 draws up a state of the art of the Geospatial Web for natural hazards. The description of the MRN Web-GIS platform is the subject of the Section 17.4. Finally, the scientific and operational conclusions as well as the development prospects are formulated in the last part of the chapter.

17.2 Geospatial Tools for Natural Hazard Exposure Modeling

17.2.1 Method: A System of Models

It is important to distinguish between modeling "of the space", aimed at identifying the properties and representing the structure of a geolocalized object, and modeling "in the space", aimed at simulating the effects of space interactions on the evolution of geographical entities (an avalanche on the slopes of a mountain, a flood within a catchment's area, etc.). Modeling the natural phenomenon ("in the space") and modeling its environment ("of the space") are not separable (Guarnieri and Garbolino 2003). According to the definition of risk defined as a function of hazard and vulnerable infrastructures, the following diagram summarizes the required processes for risk modeling:

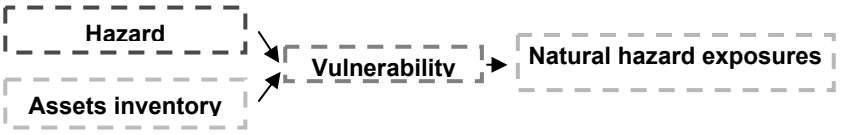


Figure 17.1: Deterministic modeling of natural risks (adapted from Kunreuther and Grossi 2005)

It imposes the mobilization of data resulting from mathematic and deductive models (such as the equation of "Barré de Saint-Venant" to model the flood flow), which must be combined with data resulting from pragmatic and heuristic models (such as the process of vector geographic information extraction from raster satellite images). The data sets can be divided into two categories: the geographic data category, which involves raster data (like digital elevation model, satellite observations of flooded areas, land cover, etc.) and vector data (like administrative boundaries, hazard maps, networks, buildings, etc.), and the alphanumeric data category, gathering preventive and regulatory information, damage statistics, vulnerability, weather data, etc. Parts of the required methods for the treatments of heterogeneous data sets are inspired by geographical science methodologies.

17.2.2 Data Requirements

Natural hazard exposure modeling can only represent a part of the complex reality. Bibliographical analysis reveals that the required data are always of the same type.

Referential data sets. The referential layer constitutes the essential input data set for the analysis of risk. It is divided into several levels from basic referential geographic data such as administrative boundaries to more contextual and more professional data such as insurance coverage conditions. Thanks to spatial analysis techniques, it supports the addition of numerous layers as well as their treatment.

Hazard data sets. The natural events are the individual or combined demonstration of natural agents from various origins having a spatial and a temporal dimension. It is important to describe the resulting data from their model because it conditions the reliability of the information. Two data types exist: (i) data stemming from empirical models, which do not describe the physical processes based on the mechanisms occurring during natural phenomenon (for example, the historical and floodplain maps); (ii) data coming from semi-physic models, which integrate physical parameters (slope, roughness and absorption of the ground, for example) aiming at producing a mathematical equation based on the creation of hazard maps (such as a rate of flow, a height of water, a velocity, etc.).

Therefore, flooded area maps allow us to determine retrospectively the costs of past events and "as if" scenarios, whereas flood-prone area maps and flood hazard maps, which both are validated and extrapolated thanks to the first ones, are useful to realize prospective scenarios. In France, local authorities are in charge of elaborating a map package relevant from these two kinds of cartographies. Finally, we should note that climate change impacts on flood frequencies and gravity have not yet been integrated.

Assets data sets. The assets are often reduced on the whole to the infrastructures and their contents, which both have an economic value (except for human lives, environmental and immaterial flows). They are generally modeled by the national geographic institute or, more and more frequently, by the local communities. They are appreciable at various geographical scales (from an aggregated scale like the land cover map to an oriented object scale like a topographic map) and supplemented for individuals by socioeconomic statistical data resulting from the population census. It is important to note that it is possible to position each asset on the territory from its address or its GPS coordinates thanks to the geocoding techniques.

Vulnerability data sets. These are alphanumeric data sets. Vulnerability is generally expressed as the level of foreseeable impacts of a phenomenon on the assets. This concept has become operational with damage functions built from statistics, expert analysis and simulations. Damage functions determine the rate-of-loss experience of exposed goods (individual or aggregated) using representative parameters of hazard intensity (generally, the height of water for the flood). It refers to tangible damage (with monetary value) for both direct and indirect damages. They can be associated to a risk category (dwelling, trade, industry, agriculture), to a value (by m^2 per land cover type, cost of replacement, average cost of the infrastructure, etc.) or to a good's insurance coverage (building, contents, trading loss; etc.; Meyer and Messner 2005). However, it should be noted that they do not integrate the response capacity of the infrastructure or the territory exposed during the crisis (Mengual 2005).

17.2.3 Geospatial Modeling Tools

The presentation of the processes that have to be implemented and the diversity of data to be handled emphasize the multifield expertise that has to be mobilized. It requires competencies in the following scientific themes: engineering sciences (hydrology, hydraulic, geology, etc), geography, geomatic, statistics, informatics, etc. The geo-informatic systems are defined both as the software techniques and the data processing for developing Geographic Information System (GIS) applications (information technologies, data warehouse, development language, GIS desktop, modeling tools, etc). They constitute adapted methodological instruments for natural risk modeling (Meyer and Messner 2005).

Geo-informatic services offer many functions of space data processing, classified according to the task scheduling inside a GIS (Bordin 2002): (i) acquisition, import, export of data sets; (ii) management and basic handling of data; (iii) request and spatial analysis; (iv) presentation and visualization.

Some private services are developed by business software companies dedicated to insurance portfolio modeling. The most famous catastrophe modeling firms are AIR Worldwide,[91] EQECAT[92] and Risk Management Solution[93] models. For more information the reader may read the publications of Professor Howard Kunreuther of the Wharton School (Kunreuther and Grossi 2005).

In summary, the evaluation process of natural hazard exposures requires collecting, integrating and organizing referential layers, natural hazard maps, and assets data sets in a geospatial platform. It is necessary to develop specific methods to produce qualitative and quantitative analysis of natural hazard exposures. The next section explains and illustrates how these mechanisms are implemented in geospatial Web environments.

17.3 Geospatial Web for Natural Hazards Exposure Assessment

Today the use of GIS is generalized for all professions in charge of natural risk assessment. Many geospatial applications have been developed for predicting, preventing and protecting against the potential damage from natural hazards. Currently, several commercial standalone desktop GIS software systems dominate the geographical information (GI) industry, such as ESRI ArcInfo and ArcView or MapInfo professional. Different vendors have their own proprietary software designs, data models and database storage structures. After having presented the geospatial risk modeling, this section defines the notion of the geospatial Web for natural hazard exposure assessment.

17.3.1 The Notion of the Geospatial Web

The development of the World Wide Web creates a unique environment for sharing geospatial data. Thus, GIS companies convert their products into "Web-enabled" tools. This evolution gives the possibility to risk management organizations to develop Web portals with Web mapping services. Many of the commercial Internet GIS programs, such as ESRI's MapObject IMS and ArcIMS (ESRI 2004) or MapInfo's MapXtreme, are developed to offer better tools for data sharing over the Web. Like the desktop GIS software, however, these Internet GIS programs also have the problems of proprietary software designs, data models and database storage structures. The sharing of data, facilitated by the advances in network technologies, is hampered by the incompatibility of the variety of data models and formats used at different sites (Choicki 1999).

Moreover, geospatial Web Services are usually not complete GI systems, in contrast to desktop GIS, which offer a broad range of functionalities. Nevertheless, data and processing tools can be integrated locally in a Web client like a standard Web browser and provide functionalities like a real desktop GIS. Different types of data and analysis tools are allocated to different servers. Such distributed GIS enjoy the advantage of the Internet as a giant distributed system (Peng and Tsou 2003).

In order to implement these GI Web services, some standards and specifications have been developed, like the International Standardization Organization's ISO 19115 and ISO 19119 (ISO 2003) and the Open Geospatial Consortium (OGC

2003). These standards define the architecture allowing the search, the access and the retrieval of geodata and GIS analysis components from any server. Some other projects like the European Commission-funded ORCHESTRA (Annoni et al. 2005a) try to improve the interoperability by implementing open services with a service-oriented architecture.

17.3.2 Overview of Available Geospatial Web for Natural Hazards Exposure Assessment

Various Web mapping services in the field of natural hazards are accessible on the Internet. A large set of services is available: consulting several data layers, downloading geographic data sets so that the end users realize their own computer data treatments, uploading and geo-positioning their own data on the interface in order to visualize their situation on hazard maps, editing reports that summarize the technical parameters of an (individual or aggregate) hazard exposure. It is important to underline that most of the current online solutions combine referential and hazard layers but do not offer a complete risk analysis.

The reference in this range of products is the recent Austrian portal,[94] which combines satellite observations and hazard layers. End users are given the possibility to position an address or GPS coordinates on the territory. It will soon be the case for France's GeoPortail,[95] where it is currently possible to visualize the French administrative boundaries, satellite observations and street maps and have access to the metadata GeoCatalogue.[96]

Currently, portals tend to improve their mapping offers as they develop data distribution functions. One of the best examples is the Federal Emergency Agency mapping information platform where citizens and insurance companies can visualize hazard maps (flood insurance rate map FIRM, hydrologic hazard maps, geologic hazard maps, atmospheric hazard maps) (Lowe 2003). The geographic data sets are not downloadable from the interface, but it is possible to import external geographic data sets for overlay layers. Users can save maps in the Web Map Context format (WMC). The service of geo-positioning is also available by keying one street address, by selecting a FEMA community or by entering GPS coordinates.

The USGS National Map Viewer[97] is a tool allowing the access to all the administrative, satellite, geologic, land cover and other topographic available maps (Kelmelis et al. 2003). Not only does the interface provide the same range of services (as previously) such as point-based querying, geo-positioning and overlaying with hazard maps (only hurricanes and wildfire maps), but it also offers to download many original data sets (such as land use, land cover or hydrographical data) except for the one concerned with hazards.

None of the previous applications presents real risk-thematic maps since they require precalculated indicators. In Europe, however, the CEDIM Risk Explorer application developed in the framework of the "Risk Map Germany" project is a good example of a risk indicator-thematic mapping tool (Müller et al. 2006). The tool enables end users to overlay hazard vulnerability and risk maps and thus facilitate their own analysis. These thematic maps are built upon two main characteristics: the data characteristics (discreta, continua, absolute, relative, etc.) and the suitable graphic compositions (graduated symbol map, filled-in isoline map, etc.). For instance, for the seismic hazard end users can visualize a filled-in isoline based on intensity values for a non-exceedance probability and a graduated symbol risk map based on estimated direct loss due to damage to residential buildings. Once again, it should be noted that none of the data presented is downloadable.

To conclude this panorama, a private British Web site proposes to generate for any address a report detailing the exposure to environmental hazards. No maps are available online, however. Some similar tools have been developed inside particular insurance markets and are enumerated in the CEA report (CEA 2005).

This brief overview points out that no integrated solution is offered to Web users to let them assess the natural hazard exposure of a territory (or one of their own interest). Although many different services exist in terms of interface conviviality, data accessibility, information and indicator restitution in respect with the technology standards, the current geospatial Web solutions are still insufficient to facilitate Web users' knowledge and to guide them towards correct decision making. In the French context of natural disasters prevention, MRN progresses towards the construction of an integrated geospatial Web platform for the general interest of the insurance profession. The next section details the services and the architecture of what is currently online for this community.

17.4 Description of the MRN Web GIS Platform

The MRN association is an interface with natural hazard stakeholders and has to communicate with the insurance companies, the insured citizens, the French public authorities and local collectivities. In such a context it appears self-evident that MRN needs communicating tools. Since 2000, MRN and Pôle Cindyniques from Ecole des Mines de Paris (its technological partner) have realized continuous IT developments on a Web platform with two main objectives. The first objective is to provide a Web content management tool to classify and organize all the textual resources (documents, publications, news, statictics and work activity reports) related to natural hazard knowledge and prevention. A specific Web content management tool has been developed (with PHP MySQL) providing flexibility to build groups of metadata forms for all published documents and information. The second objective is to evaluate Web mapping tools to analyze the situation for the municipalities facing natural hazards by following up different parameters such as the level of hazard exposure, the risk prevention plan status and the publication of disaster decrees. The first technology evaluated is MapServer (Chaze and Napoli 2004), which is the realization of a thematic atlas on flood: allowing users to visualize thematic maps built on top of the GASPAR database (Gestion Assistée des Procédures Administratives Relatives aux Risques naturels), updated every month by the French Ministry of Ecology on its prim.net Web site.[98] The second technology evaluated was JmapServer Technology developed by Kheops Technology[99] in Canada. Jmap is a solution based on Java (server and client sides) and specializes in Web spatial interactive applications within three domains: Web-based networking, spatial data warehousing and location-based services. This technology has finally been chosen because it presents some facilities regarding the integration of Web spatial applications such as the connectivity of the tool (accessibility through intranet architecture with applet and HTML browsing applications), the scalability (hundreds of concurrent users, optimized for data traffic management over global networks and for data caching), the compatibility with common GIS vector and raster formats but also the connectivity to WMS compliant data servers, the facilities to link spatial information and alphanumerical information stored in relational databases, and finally its extensibility (plug-and-play extensions like geocoding or multimedia and customization development facilities with Jmap Java SDK). Moreover, Jmap is compatible with OGC standards: Jmap- Server can query Web Map Services 1.3, integrate WMS layers and is also a WMS server accessible via the WMS protocol (and via the WFS protocol

soon). Jmap is both able to query Web Feature Services (WFS) servers and to integrate Geography Markup Language (GML) data on Web client applications. Jmap 3.0 is the current version.

Currently, the MRN Web-GIS platform runs under Jmap on two main applications: a public atlas dedicated to the flood prevention situation for municipalities exposed to this hazard (thematic mapping on flood prevention), and a private atlas dedicated to flood modeling geographic data sets for the insurance industry (digital atlas of flood zoning data).

17.4.1 Thematic Mapping on Flood Prevention

This application allows users to navigate through thematic maps about the situation of communities exposed to natural hazards. Various thematic maps show a geographical representation of statistical indicators:

- the number of catastrophe decrees published per municipality: filled-in isoline map showing the number of catastrophes at the scale of each municipality;
- the maturity of the risk prevention plan per municipality: filled-in isoline map showing plan status from prescription to approbation for more than five years;
- the presence or absence of preventive information document per municipality: filled-in isoline map showing the status of legal documents published by mayors to inform municipality inhabitants;
- the level of deductible modulation per municipality: filled-in isoline map showing the level of deductible modulation according to a specific calculation rule based on different criteria like the risk prevention plan realization and the number of past catastrophe decrees.

Navigation is possible with basic GIS components, allowing users to zoom in, zoom out, pan and print maps and also by scrolling through administrative entities (from the regional level to the departmental level and to the municipality level). It is possible to overlay the different thematic maps in order to make comparisons such as the possibility to visualize municipalities that have a lot of catastrophe decrees but without any preventive and informative documents published and that do not have an approved risk prevention plan. Users also have access to a synthesis of indicators for each municipality simply by putting the mouse over a municipality polygon: a hyperlink is clickable to access the listing of all detailed information. These thematic maps have been realized with two data sources: (i) GEOFLA spatial data (from National Institute of Geography), which contains all the geometric objects of the administrative limits (regions, departments and communities); (ii) GASPAR database (from the Ministry of Ecology), which contains all the data relative to the catastrophe decrees, the risk prevention plan and the information documents per risk and per municipality.

SQL requests from the GASPAR database are linked to spatial data by binding attributes on the JmapServer project. The interest of this application is to make a comparison between the different factors of risk at the level of the municipality. The perspective is to add other thematic maps based on other pertinent indicators. For example, the insurance companies are interested in knowing the density of the population exposed to flood for each municipality; MRN maps relevant indicators such as the density of the population and the dwellings in flood-prone areas or the damage assessment at the basin scale.

17.4.2 Digital Atlas of Flood Zoning Data

A digital atlas of flood zoning data is an application that allows end users to access all digital atlases of flood zones data sets that have been collected by MRN from public state services. As mentioned above, numerous methods exist to produce flood data sets. In addition, the atlas of flood zones is realized under the direction of departmental or regional authorities. Thus, French flood data are very heterogeneous. To overcome this difficulty, a specific data model was created, separating flooded area maps and flood-prone area maps, which can be used to make a general estimate of areas prone to flooding. The degree of danger is not always determined.

17.4.2.1 Visualization Service

The different atlases of flood zone data sets respect the standards, with two different graphic semiologies: the flooded areas ("a posteriori") and the flood prone areas ("a priori"). The flooded areas are represented by the outer envelopes with a unified blue fill color and gray border lines. The flood-prone areas are represented with various levels of envelopes according to the flood frequencies: high frequency (about 1 to 50 years' return period) is represented by a bold blue stripes' pattern, medium frequency (about 50 to 100 years' return period) is represented by a normal blue stripes' pattern, and exceptional (more than 100 years' return period). These flood envelopes are represented in a Web GIS interface, and the different layers are overlaid at different scales. At the national scale, the department polygons are visible, where the indication "carte" indicates if flood data are available. Once the users decide to zoom into one department, the flood data sets are visible and overlay with the municipality polygons. Users will see the representation of the global flooded and the floodable envelopes without any distinction of the frequency classes. Moreover, a vector data layer from Corine Land Cover (CLC) is added in order to qualify the exposure; it records 44 land cover and land use classes, which represent the major surface types across Europe. Only three classes have been represented (urban continuous zones in brown, urban discontinuous in orange, industrial and commercial zones in yellow fill color). At the intramunicipality scale, users will visualize the frequency classes for the floodable data sets. The department and municipality polygons are part of the GEOFLA spatial data sets, which represent the referential geographic data. The CLC classes are a representation of the concentration of assets in some parts of the territory. The overlay of these three layers (municipality and department polygons as referential, atlas of flood zones as hazard and CLC classes as assets) constitutes an embryo of risk visualization in terms of geospatial analysis.

17.4.2.2 Point-based Navigation Service

Different navigation processes are available: by using GIS operations (zoom in, zoom out, pan, etc.), by selecting an administrative entity (region, department or municipality) or by specifying GPS coordinates. Navigation via GPS coordinates is important because it enables end users to make the link between their internal assets and the flood data. The functionality of geo-positioning insurance portfolio elements and the possibility of overlaying them with hazard maps are fundamental aspects to evaluate which sites are prone to prevention measures and to make sure that the insured citizens are sensitive to flood hazard. A module for geo-positioning by address has been developed and is currently being validated.

17.4.2.3 Flood Data Sets Download Service

Insurance professionals need to use hazard data to evaluate the maximum probable loss for potential natural catastrophes that could happen on their portfolio. Without reliable hazard data sets it is impossible for insurance companies to evaluate their portfolios' exposure. They can therefore neither negotiate the reinsurance treaties nor determine the financial amounts for technical provisions. Thus, accessibility to the hazard data sets is an important issue in the exposure evaluation assessment.

The interface offers two ways for end users to download atlas flood zones data sets: by pointing the mouse over the flood envelopes, users will download the full, original data set with its metadata set; by right-clicking the flood envelopes, users can export online the intersection of administrative limits with the flood envelope polygons. This allows users to export data at different scales for the local and global GIS analysis of assets. All data sets are exportable in Shapefile and Mapinfo formats coupled with two metadata forms: the original metadata file and the metadata file produced by the MRN in a standardized format.

17.4.2.4 Metadata Set Service

Each flood spatial data set loaded on the platform is coupled with a metadata file provided by the regional direction of the environment in charge of the realization of the flood zones maps. The metadata files are heterogeneous because they do not necessarily follow the same metadata standard. In order to ease the use of data sets for insurance professionals, it was necessary to homogenize the metadata sets. The MRN has decided to follow the ISO 19115 metadata standard. It means that each metadata set has been normalized with the essential elements of the ISO 19115 standard: the title, the originator (regional directions of environment), the abstract, the dates of creation, acquisition, publication of the data set, the access constraint (public copyright convention), the topic category (Environment), the use constraints (use of public documents), the spatial reference system (Lambert II), the spatial resolution, the supply media (online), additional information source (modeling method; hydraulic, hydro-geomorphologic, historical event mapping), the supplier (MRN references), the data format (Shapefile and Mapinfo), the date of update of metadata, the lineage (technical indications about the construction of the data set), the spatial representation type (vector), the download link to the original metadata set file, the download link to the data sets and the associated keywords are the elements of each metadata form.

An internal metadata editor has been developed for the MRN administrator to edit and validate the metadata forms. Once the data have been validated and published, an email indicating the references of the new flood data set loaded is automatically sent to all users of the platform. Then the metadata sets are accessible via the Web mapping interface by putting the mouse over the outer envelops and via a search engine portal. With the search engine, users have the possibility to access both the list of synthetic views of the metadata sets and the detail view of each data set. For all recently loaded data sets, alert icons appear in the corresponding synthetic views.

The encoding operation in XML format (Extensible Markup Language) based on the ISO 19115 and ISO 19139 standards and the integration of some OGC Catalog Web Services are the next development steps to allow more interoperability with other spatial data infrastructures such as the GeoCatalogue.

Figure 17.2: Pointing with the GPS coordinates, downloading flood data set form and synthesis label on mouse over outer flood envelope

17.4.3 MRN Platform Architecture

The architecture of the MRN Web GIS platform was designed as a client/server model. All the components were implemented in a three-tier scheme as shown in Figure 17.3. The data tier represents the data organization and storage. It is organized in a file system containing all hazard, referential and exposure maps either in Shapefile or Mapinfo format. Nevertheless the service was designed to enable future developments towards a solution based on internal and external geographic databases (with WMS and WFS connectors). Textual and statistical data are stored in file system and databases. These data are combined and computed with referential, hazard and exposure maps to build sets of risk indicators like deductible modulation or habitation and population density exposed to flood hazard. The indicators are calculated at the scale of municipality and aggregated at the department and risk basin scales. The results are stored per municipality in a risk indicator database. The server tier consists of the main applications map server and Web server allowing the access to and the display of data according to users' spatial requests. The server implementation is based on the map service Jmap server 3.0 for setting projects (with a Web administration console) and developing templates to display maps image (Java programming with Eclipse 3.0 platform) using spatial data connectors (raster, vector, WMS, WFS), data caching, SQL binding attribute functionality (binding from risk indicators database to map objects), SQL connection to the MRN metadata database. All original metadata files are indexed in the server file system. Each metadata file download link is referenced in each metadata set stored in the MRN metadata database.

The Web server is Tomcat 5.5, which is also the servlet engine. The operating system is Sun Solaris 10 on a Unix platform. Finally, the client tier consists of a Web browser with a Java 2 run-time environment.

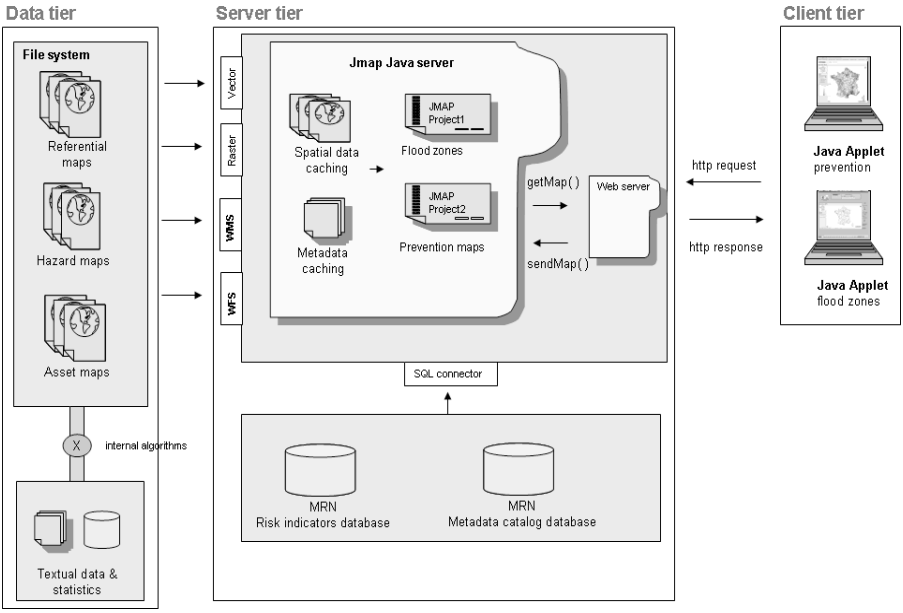


Figure 17.3: Logical architecture of the MRN Web GIS platform

17.4.4 Use Case

Actuarial sciences are the motor of risk assessment in insurance companies. In 1982, the French national authorities founded an original public-private system in order to finance catastrophic risks (natural disaster with abnormal intensity). The lack of available hazard information and damage statistics, the potential risk-adverse attitude of people and, in particular, the resulting important value of premium susceptible to being proposed to cover these risks gave birth to the constitution of a solidarity principle. Everybody pays the same rate for natural risks, which today represent 12% of any property insurance premium. Consequently, even if actuarial sciences are insufficient to model these risks in a strong uncertainty context, French insurers have never really expressed the need to seek complementary technologies in order to deal with climatic extremes.

It is the role of MRN to contribute to the expression, the structuring and the satisfaction of the French insurers' needs. The MRN Web GIS platform is the witness of the new risk engineering form that takes place in insurance companies. The requirement of the next European solvability standards but also the consequences of the possible climatic variations constitute a true challenge for the insurance industry. Two use cases are proposed in the following, and they are reminiscent of the specific French natural insurance context.

- From an insurance back office's point of view, the MRN Web GIS platform is very useful. The MRN metadata indicate which analyzes an insurer could perform with it: the probable maximum loss assessment, the expected probability curve construction, the deterministic or probabilistic scenarios development on a specific area or a portfolio, etc. The risk analyst can download hazard data sets so as to integrate them in his or her own information system. The guideline deliv-

ered with the platform shows how to present the methods, their possibilities and limits, to integrate these crucial data sets in his or her own trades in order to review the company's risk management strategy (reinsurance, alternative risk transfer, etc.).

- The second example is also quite interesting. While connecting to the MRN Web GIS platform before inspecting an industrial site, any prevention engineer will be informed of the site's exposure to natural risks. The synthetic information delivered by MRN will increase the perception level of the risk exposition, facilitating the dialog between the industry and its insurer in terms of prevention.

17.5 Conclusions

Natural hazard exposures can be represented by overlaying different geographic layers (referential, hazard, assets and vulnerability) or by mapping detailed and aggregated indicators linked to the administrative entities. Their assessment requires on the one hand organizing numerous and heterogeneous data, as well as elaborating a catalog of metadata and on the other hand providing a range of services adapted to each category of end users. The evolution of Web GIS technologies enforces risk information and risk application providers to implement new processes to communicate about hazard exposures and to provide full access to the data sets in order to serve as an input for the further processing stages. The reliability and the way information is displayed are key points to prevent the misuse of information (Gervais 2003; Devillers et al. 2003). Therefore, the architecture must take care of standards concerning the metadata structures (INSPIRE directive, ISO 1911X) to respect the integrity of the original data sources. It is also concerned with the technological standards to import and export geographic data sets (OGC, WMS, WFS).

The MRN Web-GIS platform has been online since July 2006. It provides the public with free access to the thematic mapping service. The digital atlas of flood hazard service is available only to the insurance professionals (working on portfolio exposure, maximum probable loss and prevention strategy). New services are considered for further development, such as the geo-positioning module by keying addresses with the possibility to upload a set of assets address lists and the edition of a report containing all the risk data for any point queried. At later stages other natural hazards (drought, landslides) and other digital data like digital risk prevention plans should be integrated.

As MRN is an encouraging public-private partnership model for better natural risk prevention and mitigation, it appears logical that technological interoperability between public and private spatial data infrastructures should be improved. The complexity of such a project is to combine referential, hazards and assets spatial data infrastructures involving different actors from public, private or both fields.

Chapter 18

Development, Implementation and Application of the WebGIS MossMet

Roland Pesch • Gunther Schmidt • Winfried Schröder •
Christian Aden • Lukas Kleppin • Marcel Holy

Abstract. *Since 1990, "Heavy Metals in Mosses Surveys" have been performed every five years in at least 21 European countries, including Germany, in order to map spatial and temporal trends of the metal bioaccumulation in terrestrial ecosystems. The monitoring data consist of measurement data on metal loads in ectohydrical mosses as well as site-specific metadata to characterize the sampling locations with regard to, e.g., vegetation, land use and the distance of the sites to emission sources. To optimize the data handling for the moss survey 2005/06, we developed the WebGIS MossMet with the help of open-source components. Thus, the metadata can be integrated with the information system via the Internet by the moss samplers. The WebGIS MossMet comprehensively documents the metadata, the measurement values and statistically derived metal bioaccumulation indices regionalized for ecoregions depicting the landscape coverage of Germany. In the German moss survey 2005/06, the WebGIS MossMet was applied routinely.*

18.1 Background and Goal

Assessment and monitoring of terrestrial ecosystems in Europe have a long history. Early attempts can be traced back to the 18th century, when forest inventories began to be conducted to ascertain the extent of wood resources. In recent times, monitoring activities have been predominantly initiated by emerging environmental issues such as air pollution (Parr et al. 2002). One problem associated with air pollution is the accumulation of metals in plants due to atmospheric deposition.

Since 1990, mosses have been used to monitor the atmospheric metal accumulation in Europe. Germany participates in these surveys. The monitoring results from up to 1,028 sampling sites were published as research reports (Herpin et al. 1995; Siewers and Herpin 1998; Siewers et al. 2000; Schröder et al. 2002) and peer-reviewed journal articles (Schröder and Pesch 2004a, b). The primary data collected in these surveys consist of (1) measurement data on metal accumulation in the mosses, and (2) site-specific metadata to characterize the sampling locations with regard to, e.g., vegetation, land use or elevation, and the distance of the sites to trees and emission sources (e.g., roads, motorways, human settlements or industrial plants). In previous studies, these metadata were used to assess factors influencing the metal bioaccumulation (Pesch and Schröder 2005; 2006).

In Germany the federal states are responsible for the moss collection, and the federal government finances the coordination, the chemical analysis and the reporting of the monitoring results. Therefore, in previous surveys up to 10 different moss samplers performed the moss collection, including the documentation of the sampling sites characteristics. Until 2000, all monitoring results were stored in ASCII,

MS Excel and MS Access data files. In the campaign of 2000, all these data files were integrated into the German Moss Monitoring Information System (GEMMIS) relying on commercial software products (Schröder et al. 2002).

To optimize the compilation, quality control and integrated assessment of metadata and measurement data for future moss surveys, a centralized data handling approach is needed. This chapter shows how this goal could be achieved by a WebGIS solely based on open-source products. The WebGIS should comprehensively document all monitoring results from previous surveys as well as additional geo-information on, e.g., related environmental monitoring activities and additional data like those on land use. With regard to the moss survey 2005/2006, the moss samplers should be enabled to digitize their hand-written protocols interactively via the Internet. Furthermore, the application should assist in locating the monitoring sites and controlling the compliance with the experimental protocol.

18.2 The European Metals in Mosses Surveys

Since 1990, the Metals in Mosses surveys have been carried out every five years in at least 21 European nations, including Germany (Rühling 1994; Rühling and Steinnes 1998; Buse et al. 2003). In Germany, the mosses were collected at 592 (1990), 1026 (1995) or 1028 (2000) sites (Herpin et al. 1995; Siewers and Herpin 1998; Siewers et al. 2000; Schröder et al. 2002). The monitoring data serve to map spatial and temporal trends of the metal bioaccumulation in terrestrial ecosystems according to the UNECE (2005) experimental protocol. These guidelines provide information on how to perform the sampling with regard to the moss species to collect and critical distances to keep from trees and emission sources. They help to assure the comparative ability of the measuring data. Therefore, they are indispensable for the interpretation of the spatial and temporal trends of metal bioaccumulation to check if the monitoring meets the mandatory requirements. According to the guidelines, the mosses should be sampled in autumn (September and October). The species must be collected according to the following priority: *Pleurozium schreberi (P.s.), Sclerodium purum (S.p.)* and *Hypnum cupressiforme (H.c.)*. To avoid direct influences of canopy drip by trees, the mosses should be sampled at least 5 m away from tree crowns. The mosses should be collected 1000 m from industrial plants, motorways and highly frequented roads as well as 300 m from human settlements and agricultural crop land, managed grassland and intensive animal husbandry facilities.

The German moss surveys are based on a methodically harmonized, quality controlled analytical system. After the preparation, homogenization and destruction of the moss samples, several analytical methods were used to analyze the metal concentrations (Schroeder et al. 2002). In the first three campaigns, the chemical analysis focused on arsenic (As), cadmium (Cd), chromium (Cr), copper (Cu), iron (Fe), mercury (Hg), nickel (Ni), lead (Pb), antimony (Sb), vanadium (V), titanium (Ti) and zinc (Zn).

In the first three surveys, metadata on factors that might influence the metal accumulation in mosses such as land use or near by emission sources were documented by the moss samplers according to a sampling protocol. The metadata were digitized with the help of an MS Access database application. In 2000, this database was linked with the measurement data and additional surface data in a GIS environment in order to check if the monitoring met the mandatory requirements and, if not, to prove how any deviation influences the metal bioaccumulation (Pesch and Schröder 2006). Figure 18.1 depicts the relational structure of the database consisting of one main relation ("Proben"), which is connected to 16 other relations pro-

viding detailed information about vegetation, land use, soil, administrative aspects as well as the results of the chemical analysis. In the three former surveys, 1,247 sites were sampled at least once. They were described by 2,854 metadata sets; 2,646 of them were described by concentrations of up to 40 elements.

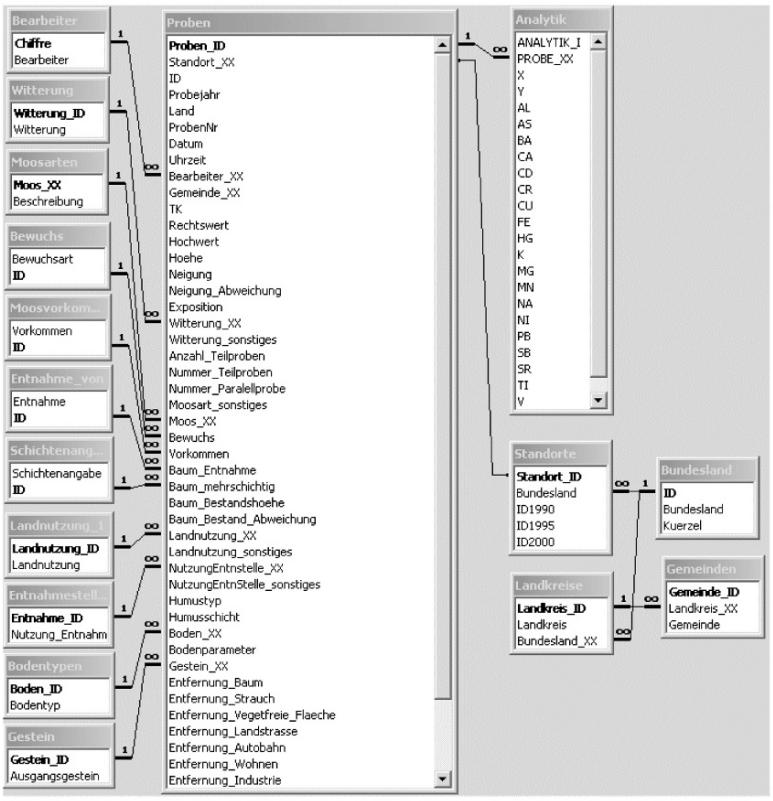


Figure 18.1: Relational structure of the moss database

Estimations were computed and mapped from the site-specific measurement data surface (Schröder et al. 2002; Pesch 2003). In further investigations, the metal-specific data were aggregated to a multimetal bioaccumulation index by means of cluster analysis. In this way regional bioaccumulation categories could be identified (Schröder and Pesch 2004a). In a further study, Schröder and Pesch (2004b) introduced an ordinal scaled index that was computed by percentile statistics ranging from 1 for low to 10 for high metal accumulation. Both indices aggregate the element loads of eight metals (Cr, Cu, Fe, Ni, Pb, Ti, V and Zn), which were measured together in each of the three surveys. The ordinal indices were intersected with a statistically calculated map depicting the landscape coverage of Germany (Schmidt 2002). In this way the metal bioaccumulation could be regionalized for areas that are quantitatively described with regard to abiotic characteristics like soil texture, climate and elevation. Temporal trends of the metal bioaccumulation can therefore easily be detected for each ecoregion. As can be seen in Figure 18.2, a continuous decrease can be observed for the three ecoregions 19, 46 and 28 from 1990 to 1995 to 2000.

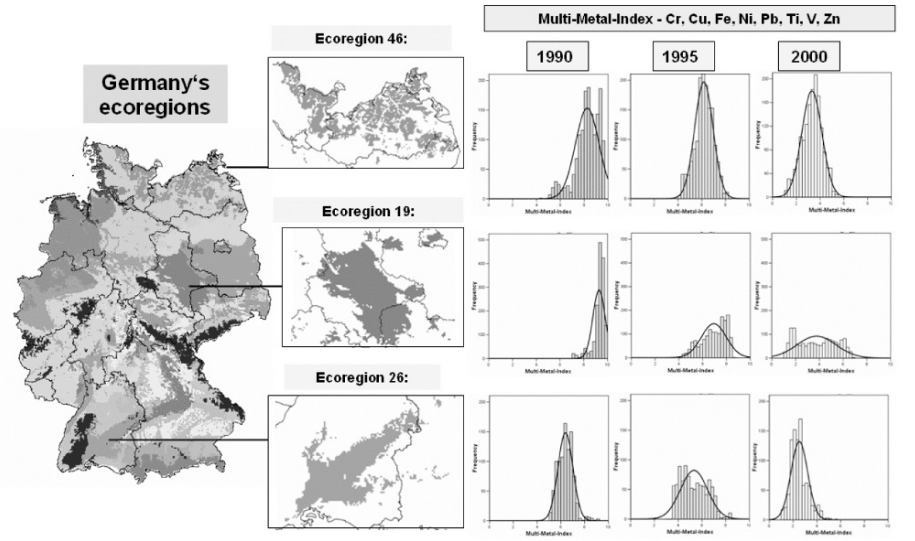


Figure 18.2: Metal bioaccumulation within ecoregions

18.3 The WebGIS MossMet

The use of Web-based information systems in natural sciences and engineering is increasing rapidly. The range of applications is wide due to the heterogeneous requirements of science and economy. Systems designed for the efficient retrieval and visualization of monitoring data and for supporting land use planning and engineering are only two examples of applications. Furthermore, such systems are appropriate vehicles for publishing and illustrating research results and to accomplish legally obligated report duties as, for example, in context of the EU-Water Framework Directive. Web-based information systems that allow spatial queries (Web mapping) can be implemented in different ways. One simple but time-consuming and traditional way is to embed static maps into HTML documents without any GIS functionality. Interaction is only possible by linking pixel coordinates with additional HTML pages or other static or dynamic contents like images or databases (clickable maps) (Dickmann 2001). To realize such systems, at least a Web server is needed as well as a database management system to handle the data flow. Such a flexible and multifunctional system is realized economically and technically at best by using platform-independent, open-source software components instead of proprietary software. Open-source products must be free of charge, and the source code is disclosed and free for modifications, in contrast to proprietary software.

The Open Geospatial Consortium (OGC) released non-proprietary standards and specifications for interfaces to process various types of geodata on the Internet. The OGC is an international organization composed of 320 businesses, universities and public utilities.[5] The amicably determined standards and specifications are supposed to ensure interoperability between various map services and to provide access to complex spatial information (Mitchell 2005).

18.3.1 System Architecture and Software Components

The software components and the system architecture of the Web GIS MossMet are illustrated in Figure 18.3. A combination of the Apache HTTP server with the WebGIS-Client Suite Mapbender, the UMN Map server and the database management system PostgreSQL including the spatial extension PostGIS was utilized.

The main function of the Apache HTTP server relies on the communication with Web clients via HTTP (Hypertext Transfer Protocol). Client-sided requests to a Web server can access documents or various file types, mobile components like Java Applets or ActiveX controls and attributes in databases or server-sided software (Peng and Tsou 2003). When the Web server receives a request for a PHP document, the source code is first sent to the interpreter of the Web server and not directly to the client browser, as is customary for HTML documents. The interpreter exclusively sends source codes translated to HTML back to the client. Thus, the underlying source code of the PHP document remains hidden.

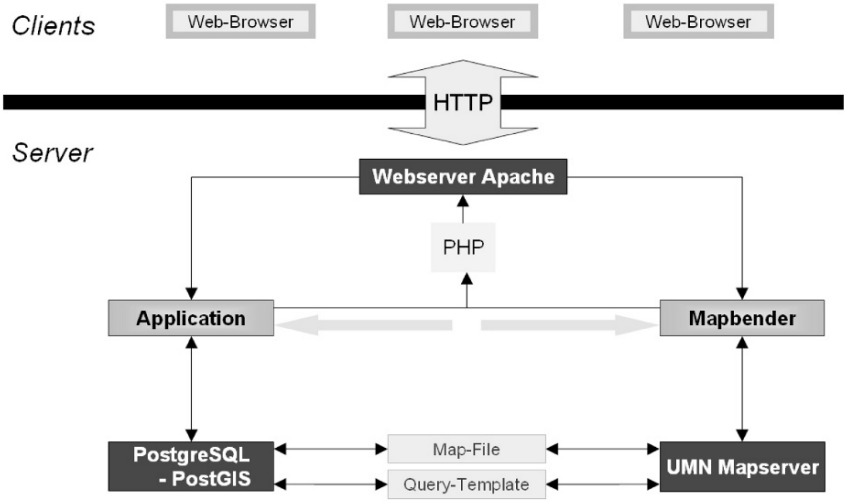


Figure 18.3: System architecture of the MossMet application (Kleppin 2006)

Map servers like the UMN Mapserver are the components that perform inquiries and analyzes on spatial data, generate maps and display symbolized maps in the form of graphics or arranged of graphic elements based on a user inquiry to a client (Peng and Tsou 2003). A map server compliant to the OGC Web Mapping Testbed specification (WMT) features different basic functions like

- generation of maps based on user inquiries,
- performance of basic queries on content and attributes of spatial objects in the displayed maps,
- extraction of data from databases,
- execution of spatial analyzes (buffering, overlaying, etc.) based on criteria from user inquiries,
- communication with other software about the existence of services and geodata.

The UMN Mapserver requires a map file to generate maps. A map file is an ASCII file containing information on paths and database connections to geodata, details on the layout of the geodata like colors or symbols, projections, scales and legends. Furthermore, the output formats as well as the metadata on the OGC-compliant operating mode of the WMS have to be defined. The WebGIS-Client Suite Mapbender by CCGIS was also utilized to generate user interfaces on the Internet. Mapbender accomplishes the OGC standards and OGC specifications on processing the geodata and offers various functions for navigation within maps, output of metadata and queries of map contents. Furthermore, it is possible to integrate remote Web Map Services (WMS).

Among the most commonly known open-source database systems are MySQL and PostgreSQL, the latter of which was used for the WebGIS MossMet. Both database systems are able to save and process spatial and geographical information in additional libraries (MyGIS, PostGIS). The Open Geospatial Consortium has published standards for types of geo-objects, functions for processing geo-objects and metadata tables (Simple Features Specification for SQL), which are supported by the PostGIS extension for PostgreSQL. Geodata that can be processed may be of vector or raster format like ESRI shape files containing polygones, lines and points as well as grids or geo-tiffs. With the help of the PostGIS extension, spatial functions and the projections are provided. PostgreSQL can be utilized for a large variety of GIS applications, including JUMP, QuantumGIS, GRASS or in connection with map servers. When data from other database applications need to be imported into PostgreSQL, Navicat PostgreSQL can be applied. This software was also used to convert and integrate the MS Access database from the previous moss campaigns referred to in Section 18.2. To integrate the metadata into the PostgreSQL database, an interface was created using PHP-embedded HTML documents. The PHP scripts contain the SQL statements needed for the data access. This part of MossMet is referred to as "application" in Figure 18.3.

18.3.2 Features

The WebGIS MossMet either digitizes new or analyzes and queries existing metadata (see Section 18.3.2.1). It is furthermore possible to visualize all monitoring data from the surveys and to relate them with geo-information on, e.g., land cover, traffic or other environmental monitoring networks (Section 18.3.2.2). Both these features are described in terms of their most important functionalities in the following. How the moss samplers may query existing metadata and furthermore digitize their handwritten sampling protocols is explained. Both applications are detailed with regard to the specifications of the sampling guideline referred to in Section 18.2.

18.3.2.1 Database

To avoid public access to the PostgreSQL database, all users have to log in before entering the WebGIS. Usernames and passwords are therefore assigned to all moss samplers and representatives of the responsible environmental authorities. The user interface of the Web database is composed of three navigation bars. Navigation bar I provides all functions of the Web application. This bar is always displayed and thus supports the operation of the WebGIS. The interface of navigation bar II depends on the function chosen in navigation bar I. Navigation bar III acts in pursuance with the function chosen in navigation bar II and enables the interactive input, output and display of all metadata and measurement data.

In order to fill out a sampling protocol for an ongoing survey, Section **A** in navigation bar I must be activated. Subsequently, the user is asked to enter the coordinates of the sampled site. When the site coordinates are entered, it can be checked whether there are sampling sites of former moss surveys within a radius of 2 km around the sampling site. The sampling guideline defines sites to be identical if they are not more than 2 km away from each other. In case the sampling site is not within a distance of 2 km to other sites, a new site ID is automatically assigned and the interactive sampling protocol can directly be accessed. If the sampling site lies within a radius of 2 km around several sampling sites of former moss surveys, one of those sites has to be selected. The site labeling of the new site is adopted from the chosen old site. If only one former site is located in a radius of 2 km around the currently sampled site, the user is automatically forwarded to the sampling protocol. The latter contains input fields for site-describing characteristics, which are significant in terms of the experimental protocol (see Figure 18.4). Input fields marked with a star are mandatory; otherwise the completion of the sampling protocol cannot be finished.

By activating Section **B** of the navigation bar I, the entire PostgreSQL database can be queried with regard to chosen criteria. All site-describing input options that are itemized in the sampling protocol can be queried and exported in a .csv file on the user's local hard drive. Furthermore, Section **C** allows changing the digitized metadata after the sampling protocol was filled out. Sections **E** and **F** give access to downloadable information like the UNECE (2005) experimental protocol, help documents for WebGIS and the addresses of the project coordination and the chemical laboratory.

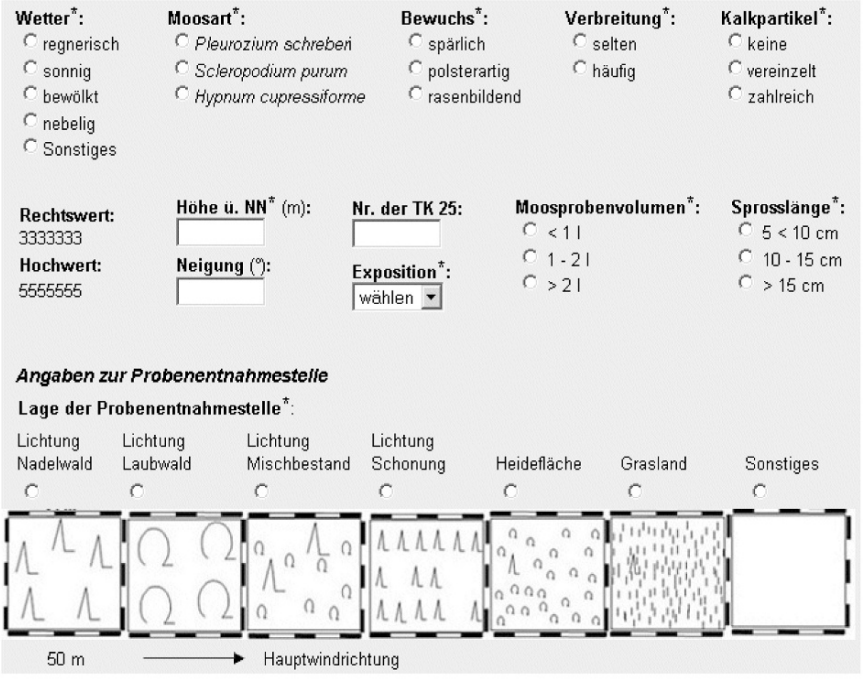


Figure 18.4: Extract of the digital sampling protocol (in German)

18.3.2.2 GIS Application

Section D of navigation bar I allows entering the WebGIS application depicted in Figure 18.4. A toolbar allows applying basic GIS techniques, for instance, automatic or interactive zooming, distance measurements or querying attribute information from selected GIS layers. A detailed map shows the respective results with regard to the selected sampling site and its surroundings, and a small scale reference map depicts the geographical location of the selected site within the whole monitoring area (Germany). The Maps & Layers tool enables the layer-oriented management of all accessible geo-objects. These include maps on administrative district boundaries, on land cover, on roads and motorways, on ecological landscape units as well as on sites of environmental monitoring networks including the moss monitoring network. In the checkboxes displayed in Figure 18.5, each layer can be visualized in the map window (left checkbox) and enabled to be queried (right checkbox). A legend is automatically generated and displayed on the right side of the map window.

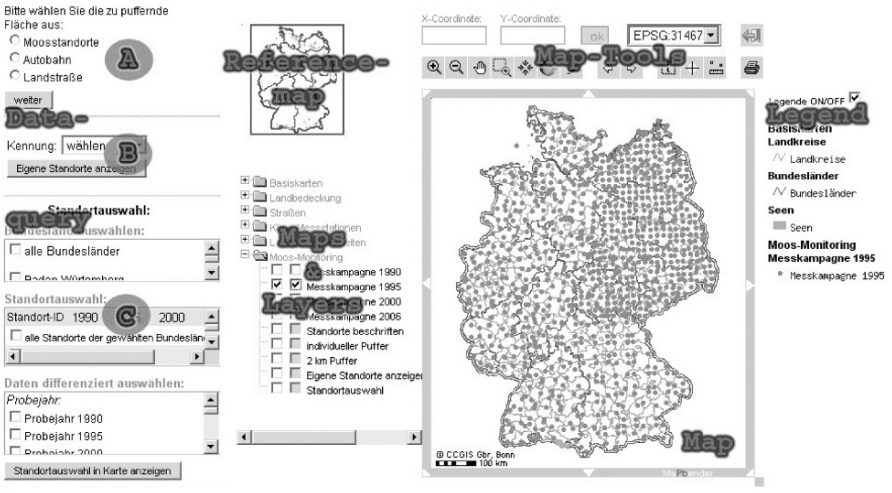


Figure 18.5: Tools and display of the WebGIS MossMet

In the area referred to as A in the data query section, the user can interactively generate buffers around the sampling sites as well as to federal roads and motorways. Instead of interactively generating buffers existent layers with predefined buffer distances can be displayed. The layer "2 km Puffer" generates a buffer of 2 km around each monitoring site. The layers "Autobahn Puffer 300 m" and "Bundesstraße Puffer 300 m" automatically generate a buffer of 300 m around all motorways and federal roads in Germany (see Figure 18.6). In this way it can automatically be checked whether the requirements of the experimental protocol were met. Furthermore, the user is enabled to assess the possibilities of where to collect the mosses when an already existing monitoring site should be sampled.

In the areas B and C of the data query section, all monitoring sites in the database can be searched by criteria of the sampling protocol. This is in accordance with the query options of the Web database referred to in Section 18.3.2.1, only here the selected monitoring can be displayed in the map view.

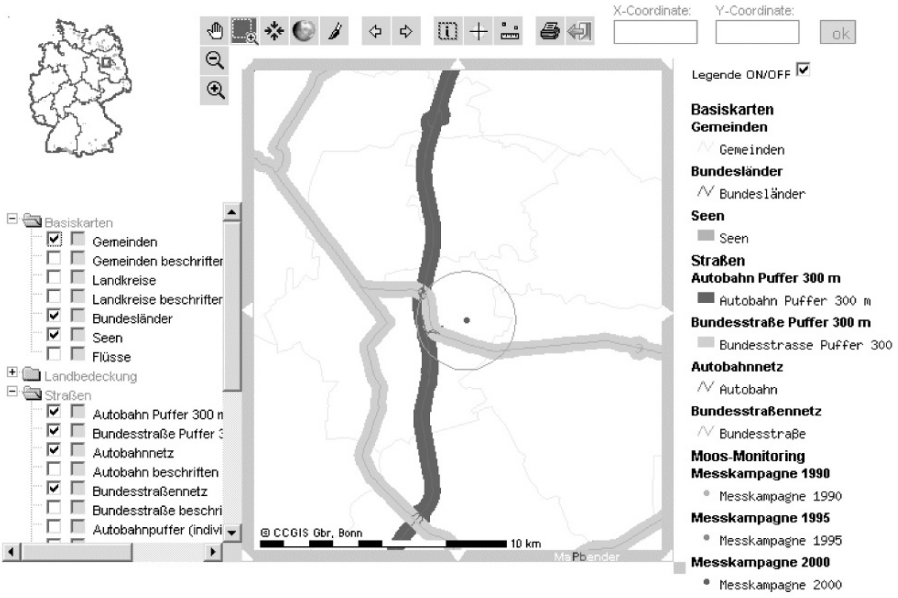


Figure 18.6: Buffering of motorways and monitoring sites in the Web GIS MossMet

18.4 Conclusions and Outlook

Environmental monitoring should provide information about environmental conditions for resource managers and policy makers and should be open to the public. To achieve a comprehensive view, data from several sources have to be compiled and integrated, or they must be collected from many monitoring sites by many collaborators. Within the framework of the European Metals in Mosses surveys, an experimental protocol should assure that the data are spatially representative, precise, accurate, consistent, comparable and reproducible. This requires techniques that help meet these quality criteria.

This chapter presents the WebGIS MossMet, which is based on the measurement data on metal accumulation in mosses and sampling site describing metadata collected in Germany 1990, 1995 and 2000. At the moment, the WebGIS was used successfully in the survey 2005/2006. Referring to users' comments on the system's applicability, we can conclude that the WebGIS MossMet effectively supports the compilation, quality control and integrated assessment of metadata and measurement data. In future surveys the moss samplers should be able to use MossMet with mobile computing devices like PDAs. In this way a mobile knowledge acquisition and checking system could be provided. Furthermore, finding adequate monitoring sites could be improved with help of the GIS-routines and the geo-information accessible through MossMet.

The differences in sampling sites and in collecting and processing of moss samples between the participating countries of the European Heavy Metals in Mosses surveys might be substantial sources of variability of analytical results. Therefore, the future UNECE Experimental Protocol should assure that information describing all 7,000 monitoring sites throughout Europe is collected and made available. The statistical analysis of such sampling site and processes describing metadata allows us

to stratify the measurement data and to compute strata-specific surface estimations on the metal accumulation across Europe. MossMet was created on behalf of the German moss monitoring program but was recommended for application on the European level at the 20th Task Force Meeting of the UNECE ICP Vegetation in Dubna (Russia) in March 2007. Up to now no other participating country has performed its moss survey with the help of Web technologies. Provided these countries are interested in using MossMet as a standard inventory tool, the application will be modified according to the differing national data acquisition procedures. The according modifications will address the database structure, design of the digital sampling protocol as well as the integration of additional geodata.

With regard to environmental monitoring issues, MossMet will strengthen the integration of monitoring sites not only within the European moss monitoring but also between different monitoring networks. This is of particular relevance because the poor integration of ecosystem monitoring at the European level causes some overlap of efforts and a lack of harmonized data. The need for a consistent pan-European long-term integrated monitoring of terrestrial systems programs is recognized in the scientific community. To reach this goal, the measurement procedures and the comparability of data are important issues, and quality assurance is a key consideration in the design process. Based on this, priority should be given to the development of compatible and interoperable databases.

Basic data on concentrations of heavy metals and nitrogen accumulated in mosses could be used for public information on the state of the environment and their changes registered by moss monitoring. This might be done by surface maps of selected parameters for different periods of time as well as by integrative maps of accumulation revealing spatial trends and structures. These findings could be complemented by a map of ecoregions that reflect the landscape structure of the study area to visualize spatial heterogeneity of deposition and accumulation. This issue should be realized by user management that allows data access on different levels of user authorization.

Additional services should be implemented to expand user interoperability in the WebGIS. These instruments should allow interactively operations, e.g., basic GIS procedures like intersecting, buffering, tabulating areas, as well as basic statistical procedures.

Chapter 19

European Air Quality Mapping through Interpolation with Application to Exposure and Impact Assessment

Peter A.M. de Smet • Jan Horálek • Bruce Denby

Abstract. *An air quality information system should offer the most complete information about air quality in a given region. AirBase contains thousands of monitoring stations across Europe, but the density varies across regions. For both public information and assessments of the exposure on human health and ecosystems, which are important indicators for air quality policy developments, the situation between stations should be known. Traditionally, assessment is based on monitoring data, but information in between stations requires accurate interpolation methods. This chapter reviews, examines and applies interpolation methodologies with special attention to the differences between urban/suburban and rural data. The methodologies are applied to ozone and PM_{10} indicators. Maps of annual average PM_{10} are shown to illustrate the recommended methodology in obtaining integrated rural- and urban-scale maps for Europe.*

19.1 Introduction

The overall aim of this study, described in detail in Denby et al. (2005) and Horálek et al. (2005), is to review, test and recommend suitable methodologies for the interpolation of regional and urban-scale monitoring data for the purposes of human health and ecosystem exposure and impact assessments, as well as for public information, on a European-wide scale. Such assessments provide important indicators on the progress in improving environmental conditions driven by policy implementations as mostly defined in both national and European Union air quality legislation. Offering such information in the form of maps on a Web-based system is expected to increase the level of providing public information and raising public awareness on air quality, as intended in current European air quality legislation (EC 1996, 1997, 1999b, 2002). In this chapter, the review and test results are presented and illustrated with accompanying maps.

The indicators chosen in this study are known as good indicators for both negative human health exposure and ecosystem effects. They are defined in EU and national air quality legislation and in UNECE protocols. They concern the ozone indicators AOT40 (EC 2002), SOMO0 and SOMO35 (Amann et al. 2006) as well as the indicators for particulate matter (PM_{10}), as annual mean and 36th-highest daily average concentrations (EC 1999a). Indicator definitions can be found in the referenced EU and WHO reports. The high-level policy attention for the impact risks of these air pollutants drives the urgent need for obtaining improved indicator assessments on these environmental pollutants and, hence, spatial information.

19.2 Review of Interpolation Methodologies

There is a significant pool of literature that addresses the question of spatial interpolation, especially its application in the geosciences. In the first part of this study, these methodologies have been reviewed (Denby et al. 2005) with particular emphasis on interpolation methods applied to air quality data. The methodologies most often encountered for spatial interpolation include inverse distance weighting (IDW), radial basis functions (RBF) and various forms of kriging. Of the studies that intercompared methodologies (Bytnerowicz et al. 2002), kriging was objectively shown to give the best results. For details concerning these interpolation methods, the reader is referred to Cressie (1993).

Spatial interpolation can be improved by the use of supplementary data. These data have a higher spatial resolution than the primary data and are in some way correlated with these. Types of supplementary data vary, dependent on the pollutant. Examples of such data used include altitude (Loibl et al. 2000), land use (Briggs et al. 2000), emission fields (Stedman et al. 2001; Lloyd and Atkinson 2004), meteorological data (Jerrett et al. 2003), satellite data (Sarigannis et al. 2003) as well as other pollutant data measured with higher resolutions (Falke and Husar 1998b). For the case of urban regions as a whole, statistical relationships between concentrations and populations have also been found as in the Auto-Oil II Program: Air Quality Report (2000). The use of model fields as supplementary data is usually approached from the modeling perspective of data assimilation, but examples of where model fields are used as a basis for interpolation also exist (Blond et al. 2003). Supplementary data are included principally in two ways, either through linear regression models (Briggs et al. 2000) or through co-kriging (Bytnerowicz et al. 2002). Other examples may use a combination of these two methods including kriging of the residual (Lloyd and Atkinson 2004). In addition to the direct question of interpolation, a number of authors acknowledge that some pollutants have a decidedly local nature and that the extent of influence or area of representativeness of an observation is limited (Beier and Doppelfeld 1999; Falke and Husar 1998a). This is the case where local emissions, e.g., in urban regions, enhance local concentrations but do not significantly affect nearby rural observations. Any interpolation method that covers both urban and rural scales should take account of this.

There are several approaches of how to evaluate the quality of interpolation. The most often used measure of interpolation quality is cross validation. This approach is empirical and objective. The values predicted by cross validation are compared to measurements in several ways: on the one hand, by linear regression; on the other hand, by forming residuals and integrating them into the statistical indicators. The most common indicator is the root mean-squared (cross-validated) error (RMSE). A smaller RMSE generally means a better estimation. Though other estimators exist, RMSE is used throughout this study.

There is currently no standard method for interpolating monitoring data for air quality assessment on any scale. It is important to identify suitable supplementary data and include them correctly in the interpolation. This means that their use for particular air pollutants and at particular scales should be tested on a case-by-case basis in order to achieve the optimum interpolation method. It is therefore recommended that such tests be carried out to achieve the particular aim of a given study.

There is no clear or tested methodology for combining urban- and rural-scale observations in the same interpolation. The need to do so is reflected in the typical spatial representation of the pollutants. This leads to the need for a more local definition of the urban contribution to pollutant concentrations.

Based on these review conclusions, a set of interpolation methodologies with their specific monitoring, modeling and supplementary input data has been selected and tested in the second part of this study (Horálek et al. 2005). The different nature of urban/suburban and rural air quality is examined as well, including a method of combining their interpolation results into one European map.

19.3 Input Data

A large set of input data is used for the interpolation methods tested. This data set includes the following sources:

- *Air quality monitoring data* that have been extracted from the *AirBase*[100] database for the years 2000–2003. For ozone, 440 rural background and 830 urban/suburban background stations were used, and for PM_{10}, 205 rural background and 724 urban/suburban background stations were used. Only stations with yearly temporal data coverage of at least 75 percent were extracted. The use of data from this European database was an essential condition for this study.
- *Unified EMEP modeling data*. Annual statistics for the years 2000–2003 have been extracted at 50 × 50 km resolution for ozone and PM_{10} (Fagerli et al. 2004).
- *Supplementary data* including altitude 30 × 30 seconds grid GTOPO30 (ESRI 2005); climatologic parameters of averaged values over period 1960–1990, 10 × 10 minutes grid from CRU (New et al. 2002): temperature, precipitation, sunshine duration, wind speed and relative humidity; CORINE Land Cover 2000 database (CLC 2000), 250 × 250 m grid (EEA 2005); population density disaggregated with CLC2000, 100 × 100 m grid (JRC 2005) for information on urbanized areas for which (sub)urban monitoring data are lacking.

19.4 Interpolation for Rural and Urban Areas

As the review recommends, the difference in nature of rural and (sub)urban air quality should be taken into account. In this regard, AirBase rural and (sub)urban background stations were assessed for the pollutant indicators of ozone and PM_{10}. The results confirm that for ozone rural concentration fields are higher than (sub)urban fields and for PM_{10} that (sub)urban concentration fields are higher than rural fields. These empirical findings can be explained by physical chemistry characteristics of the pollutants, in the case of ozone, and the local nature of emissions, for the case of PM_{10}. Due to the different nature of rural and urban air quality, it is not appropriate to use one common interpolation method simply on both types of areas together. It is necessary to create separate rural and urban interpolation maps and to produce final European maps by merging them. This is of particular importance for population exposure assessments based on interpolated data.

19.4.1 Rural Interpolation

The methodologies selected after the review process include regression models (fits), kriging, co-kriging, IDW and combinations of these using supplementary and model data. Interpolation of residuals, i.e., the differences between a model, or regression model, data fields and monitoring data is also investigated.

Each of the interpolation methods listed below, with method numbering of Table 19.1, is analyzed in terms of the RMSE derived by cross-validation analysis. RMSE is used to define the quality of the interpolation.

- pure interpolation methods using monitoring data only: (1a) IDW; (1b) ordinary kriging (OK); (1c) ordinary co-kriging (OC) with supplementary data; (1d) log-normal kriging (LK); (1e) lognormal co-kriging (LC);
- EMEP model: (2a) unfitted; (2b) regression fitted to monitoring data;
- interpolation using monitoring and modeling data: (3A) subtraction of the unfitted EMEP (2a) model from measurements with interpolation of the resulting residuals using IDW, OK, and OC with supplementary data; (3B) subtraction of the fitted EMEP model (2b) from measurements with interpolation of the resulting residuals using IDW, OK, and OC with supplementary data;
- subtraction of the fitted regression model, including both model and supplementary data, from measurements with interpolation of the resulting residuals using IDW (4a) and OK (4b).

The supplementary data used in co-kriging and in regression calculations include altitude and 30-year means of sunshine duration, relative humidity, temperature, precipitation and wind speed.

19.4.1.1 Test Results

The objective of the tests is to establish the preferred method of interpolating rural monitoring data with regard to ozone and PM_{10} indicators at the European level. Horálek et al. (2005) describe and conclude the test results in detail. Table 19.1 summarizes the test results for the PM_{10} annual average, the indicator used in this chapter as illustration of study results. The last two columns show the four-year average test results ranked with the best (i.e., lowest) RMSE results per method and per group marked as 1. Lognormal interpolation is only carried out with the pure interpolation methods of group 1. They show slightly better results than kriging without logarithmic transformation.

Table 19.1: Comparison of different interpolation methods for annual average PM_{10} concentrations showing the average cross validation RMSE (in $\mu g/m^3$), averaged over the years 2000–2003

No.	Method	Average RMSE	Ranking	
1a	IDW	6.59	5	2
1b	OK	6.29	4	
1c	OC with altitude	5.43	2	
1d	LK	6.15	3	
1e	LC	5.19	1	
2a	Unfitted EMEP model	12.78	2	5
2b	Fitted EMEP model	6.96	1	
3Aa	Combined with EMEP – unfitted, IDW	5.91	3	4
3Ab	Comb. EMEP – unfitted, OK	5.86	2	
3Ac	Comb. EMEP – unfitted, OC (with alt.)	5.48	1	
3Ba	Comb. EMEP – fitted, IDW	5.72	3	3
3Bb	Comb. EMEP – fitted, OK	5.59	2	
3Bc	Comb. EMEP – fitted, OC (with alt.)	5.37	1	
4a	Comb. EMEP + suppl. data, IDW	4.96	2	1
4b	Comb. EMEP + suppl. data, OK	4.84	1	

19.4.1.2 Interpretation

Based on the ozone and PM_{10} indicators for rural areas, the best interpolation method is the linear regression that includes both supplementary and modeling data in combination with ordinary kriging of the residual field (method 4b). When model concentration fields are not available (group 1), the interpolation tested as best for ozone is co-kriging with the relevant supplementary data being altitude and sunshine duration (method 1c). For PM_{10} lognormal co-kriging using the same supplementary data (method 1e) is found to be the best. When no supplementary data are available for interpolation, ordinary kriging (method 1b) is always found to be slightly superior to IDW (method 1a).

19.4.2 Urban/Suburban Interpolation

Since many (sub)urban regions are not monitored, any interpolation scheme developed must be applicable to these areas as well as to those that are monitored. For this reason methodologies that can use regression models, based on supplementary data, and interpolation schemes must be applicable on a European-wide basis and not be limited to areas around local measurements.

Regression relationships between measured concentrations of AirBase and the various supplementary data sets (population density, 30-year climatic parameters) are tested. Residual interpolation, using the regression models as basis, using IDW and OK are also tested to the urban stations to produce European-wide urban concentration fields. These interpolations are based on direct interpolation of the (sub)urban monitoring data and on the interpolation of the urban Delta, that being the difference between the (sub)urban monitoring data and the interpolated rural concentration field. Testing of the methodologies was carried out in a similar way to the rural interpolation studies. The analyzes carried out, with method numbering of Table 19.2, include:

- regression relationships between measured (sub)urban background concentrations and population density (1a), other supplementary data (1b) and rural background concentrations (1c),
- European-wide urban concentration fields based on supplementary data and residual interpolation using IDW (2a) and OK (2b),
- European-wide urban Delta concentration fields based on rural background and residual interpolation using IDW (3a) and OK (3b).

Table 19.2: Comparison of the different interpolation methods showing RMSE (in $\mu g/m^3$) for the annual average PM_{10} concentrations in 2003

No.	Method	2003	Ranking	
1a	Linear regression with population density	10.20	3	3
1b	Linear regression with supplementary data	8.48	2	
1c	Linear regression with rural background concentration field	8.33	1	
2a	Interpolation of concentrations, IDW	7.31	2	2
2b	Interpolation of concentrations, OK	7.12	1	
3a	Interpolation of Delta concentrations, IDW	7.15	2	1
3b	Interpolation of Delta concentrations, OK	7.09	1	

19.4.2.1 Test Results

The results of different urban interpolation methodologies are illustrated in Table 19.2 for PM_{10} annual average only. Only the year 2003 is shown as the other years gave comparable RMSE values with the same ranking.

19.4.2.2 Interpretation

The overall test results of the different urban interpolation methodologies show that the best urban concentration fields are created by interpolation of the urban Delta outside-city borders and by addition of the rural background concentrations field (method 3b). The best method based purely on regression, i.e., no interpolation, is the linear regression of the urban background fields (method 1c). The worst method is based on the linear regression of the population density fields (method 1a).

19.4.3 Combining Rural and Urban Interpolations

The interpolated urban Delta field and the rural field are superimposed and combined by applying a weighting function. The weighting function depends on population density such that when population density decreases to rural levels, the weighting function will approach zero, in our case less than 100 inhabitants per km^2 the rural interpolation result is used. For typical urban population densities, i.e., greater than 500 inhabitants per km^2, the weighting field approaches unity, which means the urban interpolation result is used. The function is linear between these two limits. This method is chosen based on an analysis of AirBase data, which demonstrates a convergence of rural and urban background station measurements at population densities between 100 and 500 inhabitants per km^2.

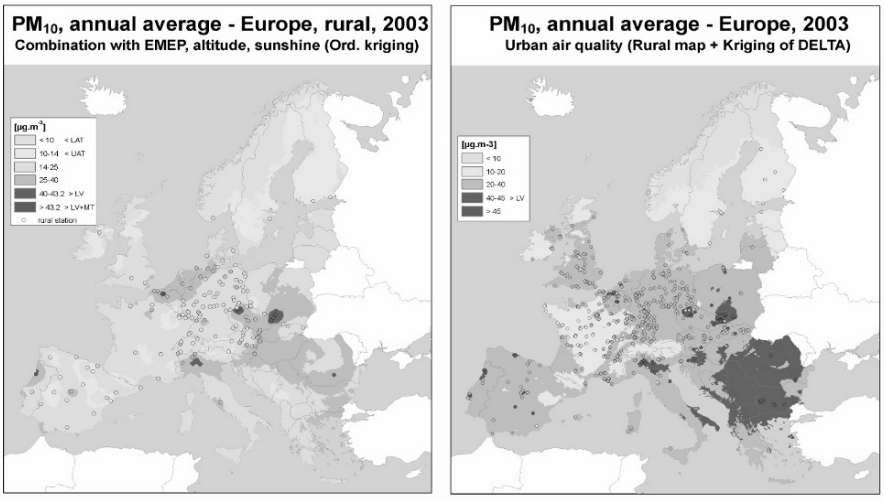


Figure 19.1: European maps for annual average PM_{10} in 2003, showing interpolated rural background concentrations fields (rural method 4b left), and the interpolated urban deltas treating Europe as if it is one large urban area (urban method 3b right)

Another reason for applying this practical methodology is that it allows for smooth merging on a European-wide scale of the rural and urban fields based on population density. It has the advantage over other methods since it does not require definition of urban borders or distances of influence, as do many other methods. It also allows the possible inclusion in the weighting function of effects that are more locally linked to regions, e.g., industrial typology, emission factors.

19.5 Application of the Recommended Method to Produce Indicator Maps for Europe

The aim of this study is to recommend suitable data and interpolation methods allowing for the most accurate regional air pollutant impact assessments in support of policy developments and Geospatial Web-based public information services. The annual average PM_{10} concentration is used here to illustrate how an integrated European indicator map is created, representing both the rural and urban areas. The following methodology is employed:

- The best rural interpolation method (method 4b of Section 19.4.1) is applied. The left map in Figure 19.1 shows the resulting rural interpolation map for annual average PM_{10}.
- The best (sub)urban interpolation method (method 3b of Section 19.4.2) is applied. The right map in Figure 19.1 illustrates this urban interpolation map.
- The urban and rural concentration fields are now combined using the weighting methodology with the population density as described in Section 19.4.3, resulting in high-resolution air quality maps of 10 × 10 km for ozone and PM_{10}. Figure 19.2 shows the final map as a result of the combined rural and urban interpolated PM_{10} annual averages for 2003.

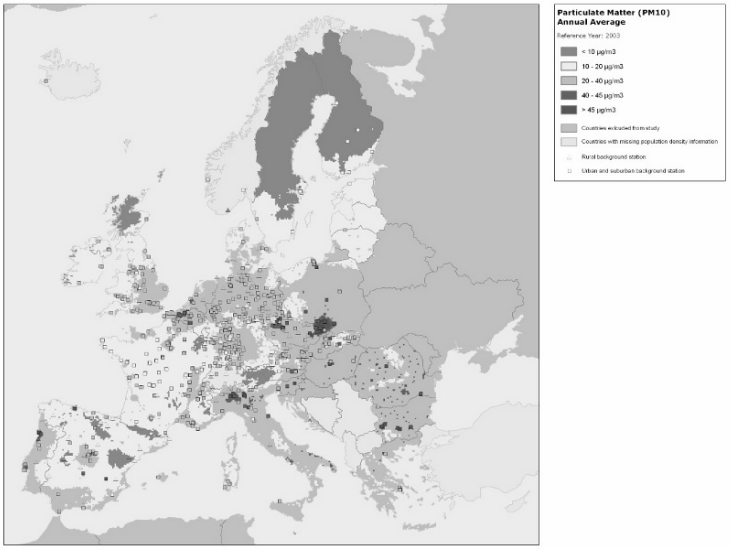


Figure 19.2: The combined rural and urban interpolated concentration map of annual averages of PM_{10} for 2003, including the measuring points

19.6 Conclusions

Based on the results and interpretations in the previous sections, a follow-up study currently explores the options for obtaining more accurate quantifications of indicators assessing the "population at risk" and the potential risk of mortality, including its spatial distribution over Europe.

Another aim is to provide better and more transparent air quality information services to the wider public, with interactive Geospatial Web portals as the most promising means of communication. This study explores ways to reach a robust, structured and accepted "foundation" for the air quality-related information ultimately provided through such portals. It supports the need for reliable and accurate spatial air quality information on both a European and more local scale.

Further to this study, other interpolation improvements will be explored, such as using concurrent meteorological data instead of the 30-year climatic data, making detailed analysis of uncertainties and exploring temporal interpolations next to the spatial dimension. An example of a first attempt of presenting ozone near real-time measurements and tentatively spatially interpolated maps can be found at the Ozone Web[101] of the European Environmental Agency.

Acknowledgements. The European Environment Agency (EEA) supported this study. We acknowledge the input from Pavel Kurfürst of the Czech Hydrometeorological Institute (CHMI) and helpful comments from Frank de Leeuw and Rob Swart of the Netherlands Environmental Assessment Agency (MNP).

Chapter 20

Introduction to Ubiquitous Cartography and Dynamic Geovisualization with Implications for Disaster and Crisis Management

Jirí Hrebícek • Milan Konecný

Abstract. *Several large-scale data and information infrastructures (SDI) are being created (INSPIRE, GMES) to support management and decision-making processes, and they are also used for solving a wide range of problems, including crisis management. These solutions require updated, precise, interoperable and integrated spatial data and information equipped with metadata. Up-to-date information, their suitable structuring and easy access to them are necessary for supporting timely and correct decision making in emergency/crisis situations. Most such information is geo-referenced. Cartographic visualization plays an important role for a user's orientation. Visualization is not an isolated element of the information transfer process; it depends on the status of source databases, decision support models, and the behavior of users. Current solutions of crisis management employ static cartographic visualizations based on prepared models of crisis situations. The chapter concentrates on ubiquitous cartography and dynamic geovisualization of real-time models and on the project "Dynamic Geovisualization in Crisis Management" undertaken at Masaryk University in the Czech Republic.*

20.1 Introduction

Geo-information and its visualized display are nowadays not only used in stationary systems; they are also becoming an integral part of mobile workstations. Mobile workstations have to be equipped with tools for processing geo-information and corresponding communication tools and channels (communication infrastructure), which are used for providing continuous connection to stationary, usually controlling, systems; mobile stations and controlling systems exchange updating information on the dynamically changing situation. The new technologies strongly influence the development of cartography and geo-informatics. There are new trends in ubiquitous mapping, such as mobile, adaptable and sensor cartography, which provide new environments for *emergency/crisis* situations.

Cartography is originally an instinctive science. In a modern approach, *mapping is understood as the ability to create a knowledge frame of an environment in space.* Development of cognitive mapping based on perception of environment and perception mediated by cartographic products, or in other words creation of maps and the use of maps, has evolved separately for centuries (Wood 2003).

GIS have significantly influenced cartography development. Today GIS support a wide range of spatial projects; however, creation of maps remains the independent and dominant conception. *GIS have not replaced cartography*; they have equipped it with exceptionally successful technologies providing a higher level of perfection and efficiency. Development of cognitive mapping requires provision of two-way live

connectivity. This enables flexible research utilizing many data sources. Maximum potential is reached by simultaneous implementation of dynamic, interactive and multimedia tools, implicit in a model called Interactive Space Human-Map (see Figure 20.1). Nowadays, intelligent access to databases and interactive user support can be used not only for location of suitable maps on the Internet, but also for map creation and modification according to specific individual requirements. Instead of just using maps that were created by someone else in advance, these new research technologies allow individuals to use cartography interactively, based on individual user requirements, to study and present spatial information. New technologies allow a "live connection" between the instinctive inner sphere of our cognition and – via direct interaction – new generation of cartographic visualizations and thereby also with the almost infinite resources of the Internet.

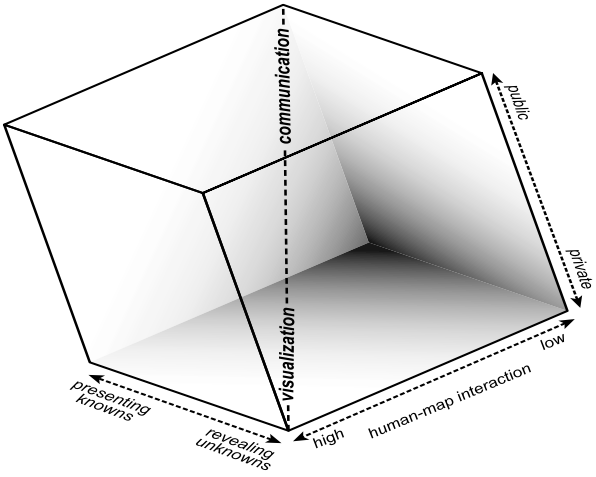


Figure 20.1: Forms of explorative visualization (modified from MacEachren 1995)

Visualization is creation of a visual image, mental or physical, with the use of graphics, photography or other tools. *Cartographic visualization* is a set of map-related graphic procedures for analysis of geospatial data and information. For example, an animated interactive digital model of terrain is a form of cartographic visualization. Cartographic visualization provides for a wide range of interpretation and representation possibilities. If they are applied on varied reference data related to the Earth's surface, we use the term *geovisualization.* Together with new cartographic GIS tools based on interactive Web maps, it is possible to mediate the mutual influence of a map and its user. Without the limitations of a fixed content and used symbols or conventional visualization techniques, the user can now create a range of maps or map-related images (e.g., profiles, terrain models) from various databases.

The map is more than just a space for manipulation and creation of images; *it can serve as a graphic window* with indefinite possibilities. This new visual-mental environment utilizes advantages of our cognitive instinctive mapping, which can be even more effective through *geovisual dialog* with cartographic/geographic visualization systems. This approach is referred to as the *cartographic revolution*, because it allows dynamic and interactive convergence of our instinctive cognitive abilities with the surrounding world of "geospatial data" (Wood 2003).

20.2 Geo-Information Infrastructures

Attempts to interconnect large amounts of distributed data have led to the idea of creating *Spatial Data Infrastructures* (SDI). Perhaps the best-known definition of these infrastructures comes from an early executive order of the former U.S. President William J. Clinton (Clinton 2004):

"Geographic information is critical to promote economic development, improve our stewardship of natural resources, and protect the environment. Modern technology now permits improved acquisition, distribution, and utilization of geographic (or geospatial) data and mapping. ... The executive branch [should] develop, in cooperation with state, local, and tribal governments, and the private sector, a coordinated National Spatial Data Infrastructure (NSDI) to support public and private sector applications of geospatial data in such areas as transportation, community development, agriculture, emergency response, environmental management, and information technology".

NSDI contains – apart from actual data – technologies, political, economic and organizational policies, standards and human resources necessary for collection, processing, storage, distribution and improvement of use of geospatial data. NSDI was the first approach to create a comprehensive SDI, but looking back we have to say that it is mostly a collection of data-set catalogs and not really an integrated data infrastructure providing immediate access to data. It also does not really provide extensibility – i.e., the ability to easily add new open data providers.

Other activities have followed in other parts of the world, including in Europe, where efforts to create a European SDI have so far not been successful. The latest initiatives of the European Union, *GMES* (Global Monitoring for Environment and Security) and especially *INSPIRE* (Infrastructure for Spatial Information in Europe), are based on three requirements: *data must be available, accessible, and follow corresponding legal conditions*. The vision of INSPIRE is to build a distributed network of databases on local, national and European levels; each database will be managed in such a way that it will provide information and services required both by individual countries and the European Union. The databases will respect common standards and protocols, providing interoperability and compatibility.

20.3 Ubiquitous Mapping

Digital cartography is strongly influenced by *Information and Communication Technologies* (ICT) and plays an important role in the *Information/Knowledge Society* (see Figure 20.2). New fields of cartography facilitate the shift from traditional maps to *ubiquitous mapping*. To solve a certain problem means to define it and strategically plan how to solve it and how to derive a solution. The research agenda comprises (Morita 2004) the following: generation of personalized maps according to the objective and spatial context; mapping system development considering participation, collaboration, and partnership of users; cross-cultural comparative studies to clarify similarities and differences between ubiquitous mapping implementations (consider information security and privacy). Morita also adds that ubiquitous mapping aims to realize technical solutions for map creation and use, and to predict the effect on society. Ubiquitous mapping accelerates, facilitates and stimulates the universal nature of map creation through the application of advanced information technologies.

One important direction of contemporary cartography is *the consideration of user demands,* allowing for real-time change of symbolic and map content. In contrast to traditional maps, maps today are produced for unique users, unique situations, and with the proper amount of information.

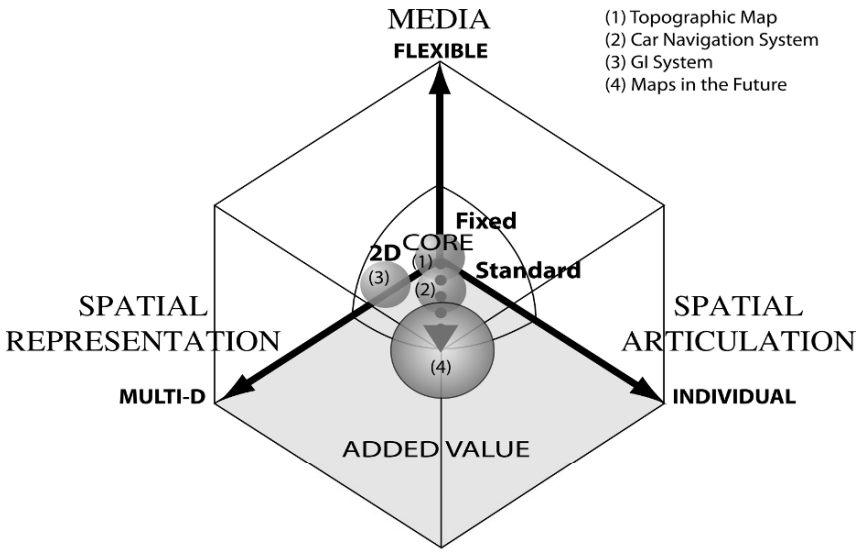


Figure 20.2: Contemporary mapping world (modified from Morita 2004)

20.3.1 Mobile Cartography

Mobile cartography is enhancing the technological part of the realization of the cartographical ideas mainly on small displays. A promising area that could utilize this approach is the area of spatial decision making concerning specialists from different disciplines and with different educational and cultural backgrounds. Geocollaboration allows various users to exchange information through common maps.

20.3.2 Sensor Cartography

Sensor cartography intends to handle and elaborate specific data and information coming from various sensors (e.g., installed along the roads or in cars or on aircrafts), their transformation and integration with data and information prepared in databases. The target is real-time map derivation for users, e.g., decision makers on a different level of public administration. Such users have different request for exactness, thematic or newly created information (Friedmannová et al. 2006) used for this kind of mapping activities through sensor term pervasive thematic mapping, which could be realized on small displays.

20.3.3 Adaptable Cartography

In comparison with the past, contemporary cartography has changed in many aspects. According to Friedmannová et al. (2006), adaptable (adaptive) cartography is one of the most important directions in contemporary cartography research. A crisis management belongs to the typical areas of adaptive cartography usage. Basically, adaptable cartography offers users more or less the same functionality as a GIS map interface. The difference is in automated processing of cartographic visualization. With GIS, cartographic visualization is user-driven – the user, according to purpose, selects map content, rarely makes generalizations and attaches appropriate symbols.

Consequent visualization is usually for that user's use only. The idea behind adaptable cartography is to automatically make proper visualization of geodata according to situation, purpose and user's background. Adaptable maps are still supposed to be maps, i.e., correct, well readable, visual medium for spatial information transmission. All map modification processes are incorporated in electronic map logic. Users can affect adaptive map just indirectly by a context. The context is the composition of characteristics describing

- *Who will read the map?* – information about map reading skills and abilities of the user, his/her visual preferences, level of knowledge and education. Such information is a base for a so-called profile.
- *Why was the map made?* – information about the task, spatial extents of the area of interest and information about map feature hierarchy according to the task.
- *Where do we use the map?* – information about place, time, orientation and environmental conditions of the map perception (for example, light conditions).
- *What is a map device?* – set of constrains influenced by display size and parameters, transfer rate from geodata source and software abilities.

The purpose of the context handling is to decrease time necessary for decision making. In the case of spatial oriented issues, the map is the natural medium for information storage and exchange. Efficiency of this information processing strongly depends on map use skills and also on *ontological homogeneity* of users' point of view. In real situations it is necessary to count with high heterogeneity the number of users collaborating on spatial related tasks. Consequently, a special map representation for every user is needed. Because it is not technically possible to create individual maps for every person in any situation, it is more feasible to create several user groups. Also, situations are divided into a certain amount of scenarios, covering the most common context combinations. *The reaction on the context change is change of the map content*. Changes are related to the particular context attributes and distinguish the following cases:

- *Change of symbolism* – the simplest and the most common method of adaptation. The change is related to display capabilities, environmental conditions and users' background or preferences. Typical implementation approach is creating symbols thesauri covering various user groups and devices. Some examples are done in the Homeland Security Working Group[102] based on OGC or in ISO standardization efforts.
- *Cartographic generalization* – quite complicated and time-consuming issue. Generalization processes react on a change of purpose, changes of features significance, changes of the spatial extent and partially also on data transmission rate. Usually, the amount of the features and features classes is reduced, and also feature representation is simplified.
- *Change of cartographic method* – is related to the users' background or to the purpose of the map. For users who are unskilled in map reading, it is profitable usage of less abstract and easier-to-interpret methods. In many cases it is impossible to separate all three types of changes.

Necessary change is usually a combination of all methods. For example, if there is a requirement for presentation of highly specialized theme to the public, it is imperative to adjust all aspects – simplify or even radically change symbolism, reduce content and finally use unequivocal cartographic method.

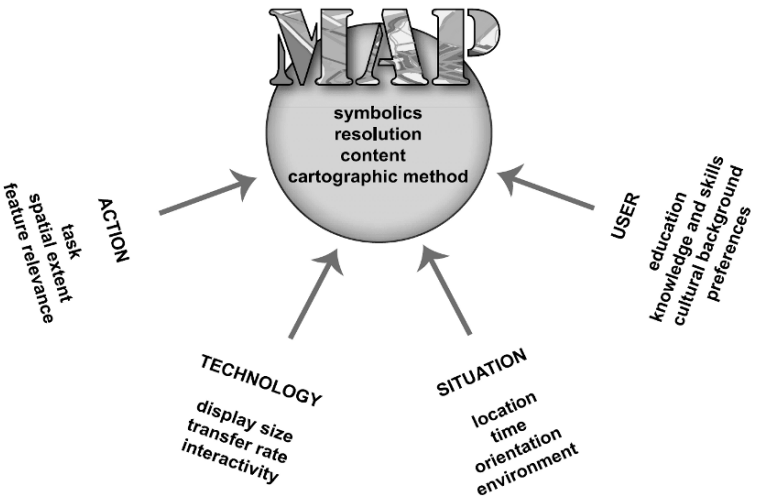


Figure 20.3: Visualization changes according to context (Friedmannová et al. 2006)

20.4 Conclusions

As Konecný and Bandrova (2006) point out, many questions asked during management of an emergency/crisis situation begin with the word "where": *where* did something happen, *where* are the rescue units, *where* are the sources of danger, *where* should the threatened people be relocated, etc. It is clear that a *natural answer to these questions is a map*. The role of cartography in crisis management is therefore clear – simplify and arrange required spatial data well. That makes the decision-making process quicker and better and leads to minimization of damage.

GIS are often the core of the entire management systems that not just solves basic localization tasks, but can also be used for planning and solving complex crisis scenarios and applying their results into practice. An integral part is also visualization of all used information both in static and in dynamic modes and also transfer and processing of all updating information.

From the point of view of geo-informatics, emergency/crisis management units utilize both SDIs including systems for collection, processing, storing, and transfer of updating, usually dynamically changing data, and methods of cartographic visualization, which communicate data and information to users' consciousness. Users' decisions – especially of those in mobile workstations operating directly in the field – are based on *visual perception* of the given information. This perception is influenced by each user's specific situation. The resulting effect of communicated information is determined by the following: (i) characteristics of geo-information (content, resolution, quality, timeliness, etc.); (ii) suitability of cartographic visualization for the used terminal device; (iii) quality of updating including application of stable transfer systems especially for updating data; (iv) an end user's psychological condition, given by his or her personal character and situation, and psychological condition of rescued persons.

Acknowledgements. The project "Dynamic Geovisualization in Crisis Management" (MSM 0021622418) is supported by the Ministry of Education, Youth and Sports of the Czech Republic (Konecný et al. 2005).

Chapter 21
Fire Alerts for the Geospatial Web

Graeme McFerren • Stacey Roos • Andrew Terhorst

Abstract. *The Advanced Fire Information System (AFIS) is a joint initiative between CSIR and Eskom, the South African electricity utility. AFIS infers fire occurrences from processed, remotely sensed data and triggers alarms to Eskom operators based on the proximity of fire events to Eskom's infrastructure. We intend on migrating AFIS from a narrowly focussed "black-box" application to one servicing users in multiple fire-related scenarios, enabling rapid development and deployment of new applications through concept-based queries of data and knowledge repositories. Future AFIS versions would supply highly tuned, meaningful and customized fire alerts to users based on an open framework of Geospatial Web services, ontologies and software agents. Other Geospatial Web applications may have to follow a similar path via Web services and standards-based architectures, thereby providing the foundation for the Geospatial Web.*

21.1 Introduction

In southern Africa, fire is perceived ambiguously. Tension exists between fire as a crucial process in ecosystems and fire events that threaten infrastructure and life. In both cases, spatio-temporal awareness of fire likelihood, occurrence and behavior is key to appropriate intervention.

CSIR is a South African research institution with extensive experience in the fire domain, ranging from policy work to basic research into fire and ecology. CSIR has been involved in developing the Advanced Fire Information System (AFIS), a joint initiative with Eskom, the South African electricity utility. AFIS utilizes results from remotely sensed data to infer fire occurrence and trigger alarms to Eskom operators based on the proximity of fire events to Eskom's transmission lines and towers. The application premise is that knowledge of fire occurrence and location can be used to prevent power supply interruptions caused by flashovers, a phenomenon of electricity arcing out of a transmission line. Certain types of fire events beneath or close to transmission lines are conducive to flashover occurrence. AFIS has achieved good success, reportedly detecting 60 percent of fires that affected power supply within its first two years of operation (Frost and Vosloo 2006).

This chapter describes the Meraka Institute (a National Research Center affiliated with CSIR) view of the future AFIS. We discuss current AFIS architecture and use cases before describing our implementation of a Web service-oriented AFIS as a precursor to an open, semantically rich, configurable, agent-based decision support framework. We describe this as a flow from geo-information dead-ends to a point where geo-knowledge is easily available, retrievable and reusable. We illustrate where ontological representation of geo-information may be beneficial in forthcoming AFIS versions, particularly referencing the Web service interface layer. Drawing upon the deep organizational knowledge of fire held by CSIR and partners, we can build a system to supply highly tuned, meaningful and customized fire information

to users. The process of semantically enriching the service-oriented architecture followed by the eventual implementation of AFIS as an ontology-driven, agent-based system could provide a pattern for similar semantic enrichment of other applications currently using closed/service-oriented GIS systems.

21.2 AFIS 1: Geo-Information Dead End

AFIS 1 typifies a complete Internet-GIS application (Chang and Park 2006). A large proportion of the application logic, data and server components exists on a single machine, dependent on the software architecture of a single vendor. It is strongly use-case bound. When distributed data are utilized, these data are poorly described and their interpretation hard-coded into the system.

AFIS 1 has simple use cases: alert users to fire events near infrastructure; archive fire events; and allow access to archives for Web-based query and retrieval of fire event information. Data from two remote sensors are used for fire detection. AFIS 1 assumes fire event existence if a "hotspot" is observed. Hotspots are pixels exhibiting higher temperatures relative to neighboring pixels. The MODIS sensor (Aqua and Terra platforms) provides hotspot detection at approximately six-hour intervals and a one hectare scale. The SEVERI sensor (Meteosat-8) provides high temporal resolution (15 minutes), but coarse spatial resolution (five hectares). Sensor-specific algorithms deployed at the ground receiving station extract hotspots. Their deployment is a black box: data from the sensors arrive, are processed and hotspots, including positional, time and other attributes, are generated into text files.

The first use case is executed – SMS alerts are generated from hotspots and sent to appropriate users. Files are sent via FTP to hotspot client machines. On one such machine, new files are parsed, and records archived into an ESRI ArcSDE database according to sensor. Records are served to the Internet via an ESRI ArcIMS Image Service. There, fires can be queried and visualized in a spatial context. System performance is good. AFIS has remained stable, serving data about thousands of fire events. Issues arise when closer inspection is made of the data generation process.

First, each point in the system relies on human interpretation – the "coder-in-the-loop" problem. For example, hotspots are generated with little metadata, certainly none that is machine processable. Second, there is a reliance on a particular software bus to deliver data. Software using ESRI MapObjects loads hotspots into ArcSDE. ArcSDE is accessed by ArcIMS for data delivery, allowing only certain clients to access the data. Adherence to open standards-based interfaces would facilitate interoperability. Third, information from the strongly use-case-driven system reaches dead-ends – it is enough to visualize hotspots in a map portrayal client or to send off an SMS message to particular users. Little further value can be extracted from AFIS by other applications, and components cannot easily be reused. Only human users can make sense of AFIS outputs.

21.3 AFIS 2: Networks of Geo-Information

AFIS 2 addresses these issues through its design as a system of loosely coupled Web services, adhering to Open Geospatial Consortium (OGC) standards, especially those concerned with Sensor Web Enablement (SWE). This OGC initiative extends the OGC Web services framework by providing additional services for integrating Web-connected sensor systems. SWE services enable discovery of abstracted sensor assets and capabilities and well-defined access to these resources for data retrieval, alert subscription and sensor tasking to control observations (Botts et al. 2006).

SWE includes draft specifications providing semantics for modeling machine-readable descriptions of data, encodings and values, improving prospects for plug-and-play sensors, data fusion, common data processing engines, automated sensor discovery and utilization of sensor data (Moodley et al. 2006). Specifications include Observations and Measurements, Transducer Modeling Language and SensorML (Botts 2005; Cox 2005). SWE provides four types of standardized Web service interfaces. The Sensor Observation Service (SOS) (Na and Priest 2006) allows users to retrieve raw or processed observations from different sources. The Sensor Alert Service (SAS) (Simonis 2006), when stable, will provide a mechanism for pushing raw or processed observations to users, based on user-specified alert/filter conditions. The Sensor Planning Service (SPS) (Simonis 2005) allows sensor tasking. The Web Notification Service (WNS) (Simonis and Wytzisk 2003) allows asynchronous communication between services and between users and services.

The main functions of AFIS 2 are to provide fire alerts, populate a hotspot archive and allow access to the archive. AFIS 2 will generate enriched fire alerts through analysis of supplementary data (see Figure 21.1). MODIS and SEVERI data are passed through sensor-dependent algorithms, producing hotspots. These hotspots are fed into an archival database, exposed to the Internet via an SOS. This allows consumers to explore hotspot history and attributes in greater detail using the visualization client. Our custom Fire Alerting Service (FAS) pulls data from the SOS as necessary. Consumers interested in hotspot alerts register directly with the FAS, or via the client, providing subscription parameters including areas of interest.

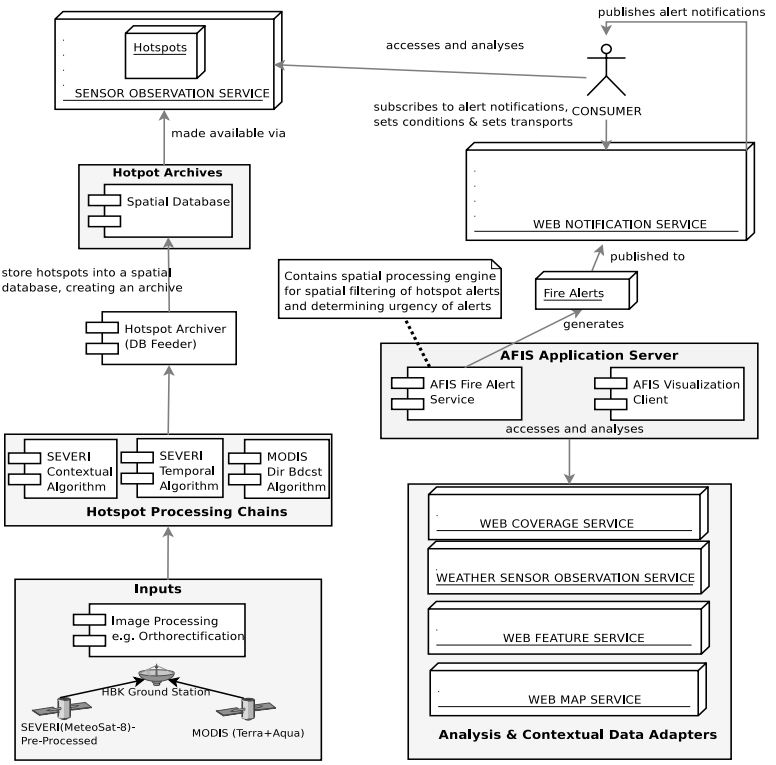


Figure 21.1: Current AFIS SWE system architecture

The FAS performs contextual and spatial analysis on the hotspots, attaching attributes to them for filtering and information purposes. An Eskom employee responsible for a set of transmission lines may be interested in hotspots within 5000 m of a line within his jurisdiction. This could be extended to include fire danger index and wind direction values. The FAS consumes data from other OGC services – e.g., a Web Coverage Service of wind stress surface generated from a weather SOS, and a Web Feature Service containing features of interest. For Eskom, features are transmission lines, towers and jurisdictions. Different use cases require alternate WFS.

Migration of AFIS to a service-oriented architecture makes the current data sources available via standard SWE interfaces and ensures that the data are encoded in SWE and other OGC data model formats. This facilitates the incorporation of AFIS 2 process chains and generated data into a network of geo-information, where they could be discoverable and reusable in other fire-related scenarios. This is a departure from AFIS 1, which incorporated no concern for reuse. The service orientation of AFIS allows services incorporating improved or different algorithms for detecting hotspots to be added to the pool of usable services. AFIS 2 moves from using the Internet to serve maps, towards using the Internet to produce knowledge.

21.4 AFIS 2.x: Towards an Intelligent System

Our current implementation of the service-oriented AFIS still suffers from many of the problems of the original "black-box" application. It is narrowly focused on hotspot provision. Spatial reasoning resides in hard-coded application logic. However, the use of the SWE services (including the public registry services) will allow Internet users to dynamically discover the information offerings, promoting reuse in various scenarios. The technical complexity of the client program has increased, as the visualization of historical data in a geographic context now requires at least two services (WMS and SOS). Hotspot data provided by the SOS must be layered over WMS data. This allows historical data querying and retrieval independently of the final image, removing the need for manual image interrogation and opening the door to automated geospatial reasoning.

Despite the Internet availability of information offerings of the services-based AFIS, end users still face problems discovering such offerings. Service discovery is done using one or more OGC Catalog Service (CS-W) instances containing information about registered services. The catalogs in the OWS-3 test bed include the ability to search based on the observed phenomena, as well as the GML names and descriptions used in observation offerings (Na and Priest 2006).

Semantic heterogeneity issues can arise with respect to the keywords used in observation offerings or phenomena descriptions (Lutz and Klien 2006). For example, AFIS may provide a "hotspot" offering, which would not match a query requesting "fire" information. Integrating SWE services into a semantic service framework improves matches between requester needs and provider offerings. We aim to provide a "semantically enabled AFIS", linking the descriptions of the service characteristics to the concepts and processes described by relevant domain ontologies.

A deep understanding of fire processes as they occur in southern Africa is vested in the CSIR. Construction of fire-related ontologies would leverage this and integrate, where useful, work that has been undertaken elsewhere (e.g., BACAREX – de la Asuncion et al. 2005). Ontologies would be grounded by higher ontologies, e.g., SUMO (Niles and Pease 2001) and NASA SWEET (Raskin 2006). We are researching OWL-S and WSMO (Feier and Domingue 2005) as mechanisms for the semantic enrichment of the workflows and processes of our current architecture.

We start with the Fire Alerting Service. This will incorporate at least one SAS, exposing another layer of functionality via a standard interface. We will provide a semantically enriched proxy layer to mediate client, sensor and service interaction, enhancing usability in the following ways:

- Attributes of particular fire events (including wind stress, fire danger index and intersection with features of interest) must be unambiguously described to allow interested users easily to determine the relevance of particular events highlighted by alerts.
- Different user profiles will allow clients to subscribe to alerts based on relevant conditions. Subscriptions will be presented such that non-expert users of the system could construct customized alert profiles, e.g., a local tower controller needs to be alerted to all fires in his or her area of responsibility; national controllers only require alerts representing significant threats to the national electricity infrastructure or supply chain.

Geo-semantic annotation of the service offering will be crucial to determine the relevance of the event to an alert subscriber. Semantically aware applications could then become cognizant of the service and use the service and domain ontologies to reason about fire processes and occurrences. To facilitate data reuse across multiple scenarios, hotspot information must be exposed before application-level processing occurs. Alert-based applications would likely be composed of layered SAS, each layer exposing richer attributes more specific to given use cases. The types of ontologies supporting the attribute descriptions would also become less abstract with each layer. Existing upper ontologies (SWEET, SUMO) would be used to conceptualize location, time and phenomena such as brightness temperature. Mid-level ontologies representing general properties of vegetation fires would allow us to describe that wind speed and direction, for example, play a crucial role in fire spread behavior. Domain-level ontologies would provide the knowledge detail necessary for describing attributes pertinent to particular applications. Lower-level SAS would be of less interest to end users than higher-level semantic SAS.

21.5 AFIS 3: Geo-Information on the Semantic Web

AFIS 3 will use an open, Web-resident architectural framework under development – the Sensor Web Agent Platform (SWAP) (Moodley et al. 2006). SWAP fuses Web Services with another distributed architectural paradigm – Multi Agent Systems (MAS). MAS include agents and the infrastructure that supports their interactions (Luck et al. 2005; Moodley and Kinyua 2006). SWAP proposes MASII, extending the standard architecture defined by the Foundation for Intelligent and Physical Agents (FIPA 2002). Standard FIPA MAS architecture components are user agents (representing end users), service agents (representing service providers), and a directory facilitator to register service agents for discovery by user agents. Agents communicate using the Agent Communication Language (ACL), possibly consuming non-agent services discoverable via a service registry.

Moodley and Kinyua (2006) introduced an application catalog containing descriptions of current applications in the MAS and Adaptor Agents for maintaining the system's ontology infrastructure. SWAP leverages infrastructure services provided by MASII to deliver new applications to end users, like the next-generation AFIS 3.

The SWAP abstract architecture has a three-tiered structure: a sensor layer, a coordination layer and a decision layer (Moodley et al. 2006). Sensor agents in the sensor layer encapsulate individual sensors, sensor systems and archived observations. For the AFIS, these sensor agents encapsulate the SEVERI observations. These data form input into the work flow agent (Hotspot Detector) resident in the coordination layer. This agent reasons with application ontologies to determine inputs required to satisfy user queries. AFIS 3 will use a non-contextual hotspot detection algorithm requiring information about expected pixel background temperature.

The Hotspot Detector retrieves both sensor data and expected background temperature (from a simulation agent, the Background Temperature Predictor). The Hotspot Detector tasks the tool agent (Hotspot Calculator) to identify hotspots based on observed and simulated data. Hotspots form input into the AFIS application agent in the decision layer. This layer also contains the AFIS client (user agent). Combining the AFIS application and client will allow users access to the full functionality of the AFIS system, including the ability to specify alerts.

Components of the open system are loosely coupled – sensor, simulation, tool and work flow agents will be available for reuse. Ontologies provide explicit descriptions of all components within SWAP, including work flows for combining components for use by software agents (Moodley et al. 2006).

Some research questions need addressing before SWAP can be realized. These relate to the internal model of agents, agent communications, message and message payload structure, ontology building frameworks, handling contradictory knowledge, integrating different ontologies into the agent paradigm, ontology maintenance, data fusion, dynamic configuration of process chains and agent development framework(s) (ibid.).

Further work would focus on the issues raised by Worboys (2005) and Cole and Hornsby (2005). Alerts can be seen as noteworthy events – happenings or activities requiring intervention; we wish to be able to reason with these events or occurences. The processes and agents that instigated these occurences and the results of these events are of interest to us. AFIS utility could be enhanced if we could reason about environmental conditions becoming conducive to particular kinds of fire or about likely behavior of detected fires. This could allow for better preparedness and sustained and adaptive management of unfolding fire events (Annoni et al. 2005b). We need to move beyond feature-based representation of geo-information to a situation where knowledge-enriched/contextualized geospatial events are but one of the possible types of meaningful content generated by applications such as AFIS.

Acknowledgements. AFIS is a Meraka Institute initiative originally funded by Eskom. Current AFIS and SWAP work is funded by CSIR. SWAP is the brainchild of a global research partnership, the Sensor Web Alliance.[103] Research partners include Meraka; the Institute for Geoinformatics at the University of Muenster; the School of Computer Science at the University of Kwazulu-Natal; Bagley College of Engineering at Mississippi State University; TNO Defense, Safety and Security; and 52o North.

Chapter 22

Geospatial Web Services: The Evolution of Geospatial Data Infrastructure

Athanasios Tom Kralidis

Abstract. *Geographic information is a valuable resource for applications and analysis where location of objects and events can enhance policy, land use and decision-making activities. Interoperability has been an ongoing activity of the geodata user community for decades, focusing on formats and standards. The recent popularity and adoption of the Internet and Web Services have provided a new means of interoperability for geodata, differing from previous approaches to information exchange. This chapter provides an overview of Geospatial Web Services as better methods to achieve efficient data exchange. The emerging Web 2.0 phenomenon is also discussed in the context of this approach.*

22.1 Introduction

Interoperability of geodata has been an ongoing activity of the geodata user community for decades. The 1970s saw the emergence of a growing requirement for national mapping and surveying agencies to create policies, agreements and processes for normalizing access to and application of geodata. The origins of a geodata infrastructure (GDI) in Canada emerged in the 1980s as an effective means of access to geodata (Groot and McLaughlin 2000). There has been an ongoing effort to produce standards-based specifications for discovery, evaluation, access, visualization and exploitation of geodata (GSDI 2001). The popularity of Web Services (CGDI 2005) has provided new means of interoperability for geodata, differing from previous approaches, such as static document-based access or nondigital acquisition methods.

This chapter presents Web Services approaches to interoperability as efficient methods for data exchange. It investigates whether Web Services are adequate for solving issues of interoperability within GDI. The following questions assist in addressing this investigation: how do Web Services address issues involving access, visualization, evaluation and discovery of geodata? What changes do Web Services create for organizations? How do Web Services affect GIS as a discipline? Taking into account relevant literature and technical publications, this chapter explores the potential of applying Web Services as an emerging approach to GDI, using specifications adopted by the Open Geospatial Consortium (OGC).

Over the last three decades, governments and industry have invested billions of dollars in the development of GIS systems to serve various information communities, such as forestry, marine, health, etc. (Groot and McLaughlin 2000). The mass of information collected by these organizations possesses the potential for multi-use and sharing between activities, systems and programs. Despite the decrease in the cost of computer hardware and software, geodata is still voluminous and an expensive resource to develop and maintain.

22.1.1 Web Services

Traditional file-based approaches have focused on static information retrieval. Once any change is made to the authoritative data set, the downloaded data set becomes out of date and inaccurate. Consumers require the most accurate, up-to-date and authoritative information from their providers. Providers would like to deliver accurate, up-to-date and authoritative information to clients.

Enter Web Services. A Web Service is a piece of business logic, located somewhere on the Internet, that is accessible through standard-based Internet protocols such as HTTP or SMTP (Vasudevan 2001). Web Services are analogous to publishing software code methods or functions over the Internet. Below are some of the properties of Web Services (Chappell and Jewell 2002):

- Primarily based on XML. Because XML provides platform-independent encoding, it represents a natural fit as the content model for Web Services to deliver information.
- Web Services allow implementations to be reworked without impacting clients. This is commonly referred to as "loosely coupled". Tightly coupled systems infer that if one interface changes, others communicating with it must be updated to accommodate the changes. Loosely coupled approaches allow for changes that allow more manageable systems with simpler interaction and integration.

Web Services allow for information to be exchanged over the Internet as a result of a customized request from a requestor. The requestor (or client) only receives what it asks for.

Consider the advantages of Web Services when applying to a database of a stock quotes archive. Past approaches would require a client to download the archive in its entirety in order to perform some sort of processing (e.g., provide the stock quote for stock XYZ on October 30, 1972). The Web Services approach enables a client to make a request to a Web Service. The following code exemplifies the response from a typical Web Service:

```
http://host/stockquote?stock=XYZ&date=30-10-1972
```

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<StockQuote>
        <Stock>XYZ</Stock>
        <Date>30 October 1972</Date>
        <Value currencyUnits="CAD">3.14</Value>
</StockQuote>
```

Although this is a simple example, Web Services enable content to be delivered as part of a larger infrastructure. The Web Services model is also well suited for a distributed, service-oriented architecture, or SOA (Chappell 2002). Note that file-based architectures are also scaleable; however, a Web Services approach reduces bandwidth requirements with just-in-time access. Client applications or tools never directly interact with the raw data. Web Services provide a representation of geodata.

There are a number of technical advantages in using Web Services:

- *Maintaining legacy systems*: Web Services position themselves as a modular approach, which allows for a developer to add a service on top of a particular legacy database or interface. As a result, internal processes are never affected, and information delivery can adhere to requirements using Web Services.

- *Independence from programming languages:* just as HTTP abstracts operating system specifics, it also abstracts programming language and development environments. For example, organization A may write Web Services using Java. Organization B may write applications using PHP. Using Web Services allows for organization B to communicate with organization A even if their components are developed using heterogeneous development environments. Using Web Services also allows development teams freedom to develop in their environment of choice, as well as alleviate restrictions of programming languages.

Web Services also have hybrid benefits in organizational/technical realms:

- *Reduction of data management issues*: the Web Services approach represents an on-demand form of information delivery. A client accessing data via a Web Service has the ability to request and receive only what is required for its purpose at a specific point in time. For example, provider A possesses a large-scale, densely populated watershed data set in the Canadian province of Ontario. Consumer B is working on a research project that requires watershed data in the greater Toronto area. Consumer B can, via Web Services, extract a subset of data for its area of interest; nothing less, nothing more. The Web Service performs the subsetting instead of the client. As a result, the researcher is not burdened with the bandwidth, data management and storage of the entire data set. Web Services enable clients to acquire the latest, up-to-date information from their authoritative provider.
- *Rapid application development and integration*: an application developer can design an application whose information is based on the Web Services approach. This results in rapid application development, lighter applications with no explicit data embedding and a multitude of data sources for integration. Consider displaying an application that has access to various data as Web Services, without ever directly integrating the data. All data are kept at source, and requested on demand by the client.
- *Reduced buy-in*: because Web Services facilitate standards-based integration, clients have less of a curve to integrate common approaches. For example, geospatial Web Services are nothing more than another set of tools and services in the Web development community, in the same manner as financial Web Services, and so on. In this context, numerous levels of buy-in are possible.

22.1.2 Open Geospatial Consortium (OGC)

The Open Geospatial Consortium (OGC) is a nonprofit, international, voluntary consensus standards organization specializing in geodata and Web Services. The OGC consists of over 300 organizations from government, academia, industry and others. The OGC was founded on the concept of providing open specifications at no cost to the public to acquire or implement. The OGC also leverages existing efforts from other standards organizations such as the W3C, ISO and OASIS.

Since the OGC WMS specification was published in 1999, the OGC has gained momentum, resulting in many early adopters of geospatial Web Services and interoperability. In fact, a recent survey identifies 166 WMS instances via Google (Ramsey 2004). It is evident that the numbers of servers indicate a level of maturity and popularity. OGC instances are found in Canada, the United States, Germany, Netherlands, Australia, Italy, Denmark, Czech Republic and Mexico (Ramsey 2004).

The OGC specifications are also making their presence felt in mainstream GIS, which can be attributed to industry recognition in response to organizational requirements based on the underlying benefits of interoperability and the Web Services approach. As of October 2006, 345 vendor products supported OGC specifications (OGC 2006d).

22.2 Benefits

When comparing the file-based, static approach of geodata interoperability to the Web Services approach, it is evident that the Web Services approach provides a lightweight, simple and efficient avenue, thanks to the advances of Internet technology and the development and implementation of standards-based approaches.

User-defined data access. Web Services offer finer-grained access to data, which allows a user to filter information. Examples of Web Services filters include (i) spatial – users can download data as per user-defined area of interest; (ii) aspatial – users can download data that meet their specific attribution criteria according to their requirements, i.e., return census data only where the census subdivision population is greater than 10,000; (iii) hybrid – the above-mentioned filters can be combined for further fine-grained access. As a result, Geospatial Web Services offer better control for data access, giving users what they want when they want. This introduces a change in paradigm from supply-driven to demand-driven for geodata exchange.

Data management. The nature of Geospatial Web Services provides a mechanism for users to discover, access, visualize and evaluate geodata in a dynamic fashion. Utilizing this approach, users are spared the need to harvest and manage data within their IT/IM domain. For example, NRCan's Landsat 7 Orthorectified Imagery data, offered as an OGC:WMS layer from GeoGratis, equates to about 1 terabyte of disk storage and capacity. A partner community wishing to access this layer for visualization does need to undertake the responsibility of data management for this data.

Data timeliness. Another advantage is timely response to information retrieval. Web Services lessen the amount of data transmitted over the Internet, resulting in faster response times.

Data quality. Data quality is of utmost importance to making sound decisions. Web Services do not directly affect data quality, as they provide only the transport mechanism. File-based approaches involve data management by clients who may not be mandated to keep the data in their pristine state. As a result, copies of data become outdated and provide another degree of separation from the original data. However, Web Services provide a higher probability for maintaining data quality.

Authoritative data sources. In a Web Services environment, information communities can benefit from maintaining only their domain-specific information holdings and leverage data from their partners without maintaining it.

Accurate and up-to-date data. The Web Services approach can ensure users always receive the most up-to-date information from their service provider. As data management budgets decrease as a result of the Web Services approach, custodians will have more time and resources to put towards data completeness, quality, accuracy and precision.

Web 2.0 and Geo. The recent proliferation of Web 2.0 (O'Reilly 2005) has resulted in highly interactive applications, as well as increased usage of remote resources. Web Services in Web 2.0 are prominent in platform connectivity in the form of mash-ups and significant use of XML. The concept of AJAX approaches has resulted in the emergence of the Web browser as an additional form of an applica-

tion platform. Geospatial Web Services are a natural extension of Web 2.0, aptly called "GeoWeb 2.0" (Maguire 2006). Maguire's vision of Geo- Web 2.0 involves broader clients, such as Google Earth, as well as Web services, standards and remote resources. Web 2.0 also enables a "read-write" information exchange in terms of end users being able to modify information over the Internet. Geospatial Web Services support this approach via the OGC WFS specification, which supports transactional operations. The benefit of GeoWeb 2.0 is the ability to develop on various platforms. As Web Services can be used in common Web browsers, they can also be leveraged in standalone desktop applications such as geobrowsers (Google Earth, NASA Worldwind, etc.), cellphones, GPS, etc.

22.3 Organizational Impact

A survey of OGC deployment over the Internet (Ramsey 2004) provides insight into how many publicly available OGC Web Services are available. These services are still in the minority as compared to traditional file-based services and proprietary systems and applications. Traditional approaches are still evident as the major method for geodata access and use in such Web sites as the *GeoConnections Discovery Portal* (GeoConnections 2001), the *Geospatial One-Stop* (U.S. Department of the Interior 2003) and the *FGDC Clearinghouse* (Federal Geographic Data Committee 1999).

For many organizations, these approaches represent additional resources to train staff, as well as change in the information delivery model to an SOA approach. While the long-term benefits to an organization undertaking this approach are many, the near-term issues in adopting this approach may be cost prohibitive, especially for smaller organizations.

The survey additionally illustrates that many of those who have migrated to this approach are large, federal infrastructures with large programs to facilitate such a change, although some smaller organizations have also begun to adopt this approach. In the context of CGDI, organizations such as Natural Resources Canada, Environment Canada and Agriculture and Agri-Food Canada are all examples of early adopters of the Web Services approach.

The scenarios discussed involve simple map production and sharing concepts. While these are common tasks in the geospatial community, so too are advanced geoprocessing tasks such as image classification, triangulation, intersection, point-in-polygon shortest-route algorithms (de Berg et al. 2000). These operations require advanced processing in geospatial software, as well as more complex interaction from a client. OGC:WMS implementations typically perform simple algorithms to convert real-world coordinate data to a 2D image in XY space. Compare this to more intensive geometric algorithms, such as buffering features or testing for feature intersection.

Groot and McLaughlin (2000) provide a comprehensive analysis of how the Internet has changed the landscape of GDI skill profiles. They argue that now there are (i) fewer mapping specialists, (ii) more Internet-aware GIS experts, (iii) GIS software experts augmented by application programmers for customized application development, (iv) increasing shifts from GIS experts to information systems specialists. Web Services add another layer of complexity for individuals to become familiar with. Though the proliferation of standards-based tools will ease integration for GIS specialists and developers, fundamental knowledge of Internet approaches is required for development and sustainability. This may require more collaboration with organizational information technology and infrastructure providers.

22.4 Recommendations and Further Research

The need for further research would be beneficial in applying Geospatial Web Services to more advanced processing tasks, such as line intersection, image analysis as well as dynamic linking of data over the Internet.

Semantics. While the current approaches of Geospatial Web Services provide well-defined operation syntax and schema for geodata, further research is required on interpretation of Geospatial Web Services' content models.

Uptake in organizations. Further research, such as that by Ramsey (2004), is of utmost importance into the organizational uptake of Geospatial Web Services. Are organizations ready for Web Services? How do Web Services affect information management and technology practices? How are Web Services best implemented in a given organization (bottom-up developer-driven, top-down management-policy-driven)? What are the cultural factors involved?

Profiling geospatial professionals. The concept of GDI has raised many issues relating to human resources and skill sets of those participating in GDI and has changed the need for specific skill sets in geomatics. The need for a well-rounded geomatics professional is emerging (Groot and McLaughlin 2000). Shrink-wrapped GIS software presents an ease of use to end users that may be detrimental to creating adequate human resources within GDI. The need for application programmers, who possess the knowledge and skills to develop and emulate desktop GIS software, has increased. Academic institutions may want to investigate these issues to provide individuals with more awareness in geomatics and GDI.

Web 2.0 considerations. While it is evident that Web 2.0 and Geospatial Web Services present a multitude of opportunities for applications, the very collaborative nature of Web 2.0 presents an area of concern for geodata. Issues of copyright and authoritativeness arise with the idea of modification of data online, in terms of trustworthiness of the source of update. Processes and policies must be investigated with regard to enabling participatory geography over the Web.

22.5 Conclusions

The discussion of Geospatial Web Services as a broad information technology approach illustrates benefits with regard to new, innovative development and integration. Geodata can now interact with more types of domain-specific data, including aspatial data, which present possibilities for new research and development. Organizations can benefit from this approach to more effectively manage their information assets and utilize other data.

Chapter 23

SWING – A Semantic Framework for Geospatial Services

Dumitru Roman • Eva Klien

Abstract. *The ability to represent geospatial semantics is of great importance when building geospatial applications for the Web. The Semantic Web Service (SWS) technology provides solutions for intelligent service annotation, discovery, composition and invocation in distributed environments. Deploying this technology into geospatial Web applications has the potential to enhance discovery, retrieval and integration of geographic information, as well as its reuse in various contexts. This chapter gives an overview of the SWING research framework, which is aimed at investigating the applicability of semantic technologies in the area of geospatial services. The goal is to provide a semantic framework that facilitates the employment of geospatial services to solve a specific task in geospatial decision making. In this chapter, we emphasize the motivation and the challenges for such a framework, point out the main components and highlight its potential impact.*

23.1 Introduction and Motivation

Geographic information is an integrated part of everyday life, and geospatial services are therefore an attractive field both for research and for practical purposes. Sustainable planning of infrastructure development, spatial occupation and resource consumption requires a long-term perspective and an integrated approach to land use management across Europe. Today, many data sets are made available through geospatial Web services.

Interoperability is supported by the Open Geospatial Consortium (OGC)[5] with a series of syntactic interface specifications, establishing protocols for the components exchanging geospatial information. However, challenges remain in supporting the crucial tasks of discovery and retrieval of information sources that meet the user's needs. Metadata standards for the description of geodata exist as well as catalog services to search them. But these do not consider that the conceptualizations governing the different implementations have been constrained in different ways (Burrough and Frank 1995), causing semantic heterogeneity during discovery of information sources and retrieval of information (Lutz and Klien 2006).

In such open and heterogeneous environments, semantic interoperability is crucial for searching data sources and evaluating their content. What is lacking is semantic annotation of the information sources, i.e., a formal and explicit representation of their semantics (Kuhn 2005), as well as a supportive environment for realizing semantic discovery and retrieval.

While the standardization efforts of the OGC concentrate on syntactic interoperability, the Semantic Web initiative has brought the semantic issues of information processing into perspective (Berners-Lee et al. 2001). It seems promising to adopt the developments around the Semantic Web in order to approach semantic

interoperability in Geospatial Web applications. Visions, architectures and applications of this cross fertilization of Geospatial and Semantic Web technology are cumulating in the notion of a *Geospatial Semantic Web* (Arpinar et al. 2006; Egenhofer 2002; Fonseca and Sheth 2002; Kolas et al. 2005).

Semantic Web Services (SWS) represent the combination of two technologies: Semantic Web and Web Services. They can be defined as "self-contained, self-describing, semantically marked-up software resources that can be published, discovered, composed and executed across the Web in a task driven automatic way" (Arroyo et al. 2004). By using developments like the Web Service Modeling Ontology (WSMO) (Roman et al. 2005) and the Web Service Modeling Language (WSML) (de Bruijn 2005) for the semantic annotation of geospatial services, one could utilize SWS technology such as the Web Service Modeling Execution Environment (WSMX)[104] to increase the efficiency and accuracy of discovering and integrating Geospatial Web services.

To describe a Web service semantically, a comprehensive knowledge of logic, ontologies, metadata and various specification languages is required. Thus, two major impediments for realizing the SWS vision in the area of geospatial services are (i) the lack of Web services that are semantically described, and (ii) the lack of development tools that can hide the complexity of and automate the creation of the necessary semantic markup.

In this context, we present a work-in-progress semantic framework for geospatial services, developed in the context of the *Semantic Web Services Interoperability for Geospatial Decision Making (SWING)* project,[105] which aims at deploying SWS technology in geospatial applications. In particular, the SWING framework addresses two major obstacles that must be overcome for SWS technology to be generally adopted, i.e., to reduce the complexity of creating semantic descriptions and to increase the number of semantically described services. SWING wants to provide an open, easy-to-use SWS framework of suitable ontologies and inference tools for annotation, discovery, composition and invocation of Geospatial Web Services.

The rest of this chapter is organized as follows. In Section 23.2 we present the core challenges for the combination of SWS technology and Geospatial Web applications. Section 23.3 highlights the core components of the framework, and Section 23.4 summarizes its expected impact. Section 23.5 summarizes the chapter with an outlook on future steps in the development.

23.2 Challenges for Combining SWS Technology and Geospatial Services

SWS technologies promise an automated information processing based on the ability to assess semantic interoperability within the information flow of service-based infrastructures. We first describe a scenario for service composition to introduce the components of Geospatial Web applications. We then list the core challenges for combining SWS technology with geospatial service and illustrate the benefits with examples from the scenario.

In the scenario, a decision maker wants to produce a map that shows the ratio between production and consumption of aggregates within a certain region. Aggregates are needed as building material for a variety of construction applications, and the visualization of the production-consumption rate gives valuable information, e.g., for the planning of new quarry sites. To fulfill this request, several services are needed:

- Web Feature Services (WFS)[61] that provide the underlying geodata: *Quarry WFS*: points for quarry location, authorized production per quarry; *Aggregate WFS*: polygons for administrative entities, average aggregate consumption per administrative entity; *Admin WFS*: polygons for administrative boundaries, population density per community.
- Web Processing Services (WPS)[106] that provide processing functionalities: *Consumption WPS*: to calculate population density × average consumption per administrative area; *Production-Consumption WPS*: takes features with production rate and those with consumption rate as input and returns features with production – consumption rate attribute.
- Web Map Service (WMS)[4] to visualize the results: Production-Consumption WMS portrays the results of the Production-Consumption WPS.

This scenario involves two use cases. First, a service developer has to find all services, construct the service composition and then register the composition in a catalog. This task will be supported by the development environment (see Section 23.3). The end-user application provides semantic search functionalities to search for services and service compositions registered in the catalog, as well as advertising a request that might not yet be satisfied with the registered services.

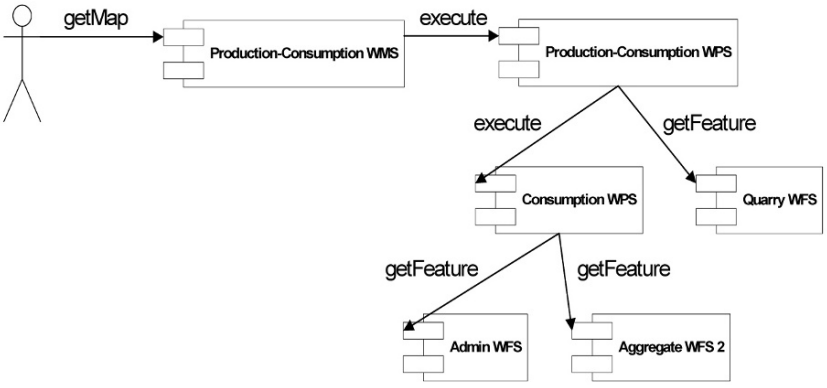


Figure 23.1: Service composition for generating production-consumption map

The strategies of the SWING framework to assess semantic interoperability focus on automated information processing: information has to be discovered, retrieved, evaluated and translated based on formalized descriptions, i.e., ontologies. In the context of geospatial Web applications, we have to capture the semantics of input and output, i.e., the semantics of data (features), as well as pre- and postconditions of operations, i.e., the semantics of geoprocessing.

To construct the service composition in Figure 23.1, the service developer first has to search for relevant services. For example, a keyword-based search for information sources that offers quarries together with information on production rates will have low precision and recall due to different terminology (e.g., pit, exploitation site, or quarry), different context (quarries for extracting stone vs. quarries for extracting slate) and the lack of complex query formulation (e.g., keyword search only considers quarries, without being able to specify the need for the information on production rates). To overcome these problems, we provide a core set of ontologies that can be used by the service providers to semantically annotate their services as

well as by the information requestors for formulating complex queries. These ontologies are represented in the logic-based Web Service Modeling Language (WSML; de Bruijn 2005), thus enabling some degree of semantic matchmaking based on logical reasoning:

- Domain ontologies should capture the specific view of an information community independent of how the information is encoded. Figure 23.2 schematically depicts parts of a domain ontology that has been developed in cooperation with the *Bureau de Recherches Géologiques et Minières* (BRGM). This domain ontology captures knowledge on quarries (aggregate productions) and construction applications (aggregate consumption).
- Ontologies that capture the specifications for geodata encoding and processing (e.g., based on OGC specifications). The explanation of a feature's encoding based on these ontologies helps to evaluate the interoperability of information sources not only on the semantic level (e.g., a feature represents Quarry) but also on the encoding level (e.g., point vs. polygon, different spatial reference systems), which is crucial for further information processing.

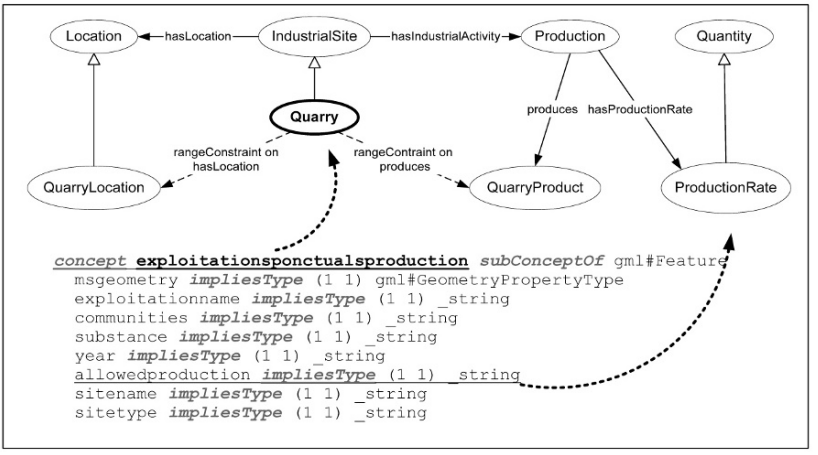


Figure 23.2: Examples of how elements of the feature type's schema can be semantically annotated with concepts in the domain ontology

Currently, metadata that are readily available for geospatial services is provided as semistructured and natural language text. To enable automated information processing, this information has to be transformed into formalized Web service descriptions. The concept definition for the feature type "exploitationponcutalsproduction" that is depicted in Figure 23.2 is the result of an automatic translation of the feature type's schema into WSML syntax. This translation process involves references to the OGC ontologies (e.g., gml#Feature, gml#GeometryPropertyType). Moreover, to ensure semantic interoperability within the environment, the feature type's description has to be referenced to existing domain ontologies. Again, Figure 23.2 depicts two examples of how elements of the feature type's schema can be semantically annotated with concepts in the domain ontology. Such a semantically annotated feature type can then be used to constrain the output in the postconditions of the QuarryWFS, i.e., a request for quarry features together with information on production rates will find a match in the QuarryWFS that serves the feature type

"exploitationponctualsproduction". *The generation of such semantic annotations in the SWING framework will be supported by a tool that automates the annotation process.* New methods need be developed and tested to provide the automatic support, e.g., analyzing geometry and topology of geodata to infer whether a feature type can be mapped to a concept of the domain ontology.

The OGC promotes service-oriented architecture along the lines of the ISO Reference Model for Open Distributed Processing (RM-ODP). Geospatial services act either on data (WFS, WPS and WMS) or on metadata (e.g., catalog services). The services have well-defined interfaces that support discovery, retrieval and execution (e.g., *getCapabilities*, *getFeature*, *execute*). But only in recent efforts have these services been aligned with the mainstream publish-bind-find paradigm represented with SOAP, WSDL and UDDI. Several peculiarities of geospatial services like the differences in service paradigm and the need for specific spatial constructs in the ontology language (Lemmens 2006) require an adaptation of existing SWS technology (WSMO/WSML/WSMX) to provide a tool to facilitating discovery and invocation of semantically described geospatial services.

23.3 Core Components

The SWING framework consists of six components, which are summarized in the following; up-to-date information on these components is available through the SWING project Web site.[105]

- The **Application** prototype, also called MiMS (Mineral Resources Management System), is the environment for the geospatial decision maker, who wants to use resources on the Web. MiMS provides the domain expert with convenient access to the functionalities offered by the other components. Queries to the Catalog, for example, can be formulated based on the domain ontology, are then automatically translated into a goal description, which makes them processable in the Semantic Discovery & Execution component.
- The **Semantic Discovery & Execution** component provides the basic SWS infrastructure. It has access to the repository of semantically described Web services and consists of a set of inference tools for discovery and dynamic invocation of the services. At the core of this component is the WSMO (Roman et al. 2005) framework – a generic framework for modeling the entities involved in handling semantically described Web services. The reference implementation that is used in SWING is WSMX along with a family of logic languages called WSML. From the WSMX perspective, SWING is an application of generic SWS techniques.
- The **Geospatial Ontology** component contains a repository of ontologies used by the Semantic Annotation, Catalog, and the Semantic Discovery & Execution components. Ontologies provide formal definitions for concepts in a domain and the terms to denote them. Ontologies are increasingly used to uniformly access information. They can be applied for making the semantics of geospatial data sources explicit and enable automated semantic matchmaking.
- The **Semantic Annotation** component utilizes information acquisition technology to analyze the semistructured data descriptions of existing geospatial Web services, in order to generate semantic annotations. Such annotations are generated based on text and data mining methods. Annotations are registered in the Catalog component in order to extend the number of semantically described Web services.

- The **Catalog** provides a standard OGC service registry interface storing entries to classical geospatial and nonspatial services. In addition, it utilizes the underlying components to provide semantically enhanced discovery functionality. Besides spatial filter and keyword-based search, the Catalog incorporates functionalities of WSMX to support semantic discovery, hence improving the quality of the result.
- The **Development Environment** component is based on IBM's Open Source Eclipse Development Environment. It consists of a number of plug-ins that integrate and hide the complexity of the other components. Application and service developers can use the Development Environment to discover both semantic and nonsemantic services, to semantically describe their services and to compose multiple services.

23.4 Expected Impact

The impact of the *SWING* framework will benefit several major user groups, i.e., geospatial decision makers, data and service providers, application developers and the research community. Geospatial decision makers will be able to use the dedicated application and hence decrease the time used to discover and utilize relevant data. Citizens are allowed to access the information from the decision-making process and can thus better understand the rationale behind the decisions. Data and service providers will be able to use the Catalog component to annotate their services. Application developers will be able to use the development environment to create semantically annotated, composed Web Services more effectively than before. The research community will be able to utilize the experience gained during the development of the SWING framework and to directly experiment with, reuse or extend the core open-source components, i.e., the ontologies, annotation engine, execution environment and the development environment.

23.5 Summary and Conclusions

The ability to represent geospatial semantics is of great importance when building geospatial applications for the Web. It will not only enhance discovery, retrieval and integration of geographic information, but it will also enable its reuse in other contexts than the original one. However, the scarcity of semantic annotation and the lack of a supportive environment for discovery and retrieval make it difficult to employ geospatial services to solve a specific task in geospatial decision making.

By deploying SWS technology in geospatial Web applications, we aim at developing a comprehensive framework for semantically annotating geospatial services and for utilizing these annotations in service discovery, composition and invocation.

Major challenges of the work rely in identifying and solving differences between OGC and SWS paradigms, to account for the specific requirements of expressing geospatial semantics, to develop intelligent methods for (semi-) automatically annotating geospatial services and in developing a pilot application that will increase the use of distributed and heterogeneous services in geospatial decision making.

Acknowledgements. This work is funded by the European Commission under the SWING project (FP6-26514). The authors would like to thank all partners involved in the SWING project for stimulating discussions around Semantic Web and geospatial services.

Chapter 24

Similarity-based Retrieval for Geospatial Semantic Web Services Specified Using the Web Service Modeling Language (WSML-Core)

Krzysztof Janowicz

Abstract. *What prevents the Geospatial Web from taking off is not a missing architecture and protocol stack but, beside other aspects, the question of how Web services can be semi-automatically discovered and whether and to what degree they satisfy user requirements. Two approaches turned out to be useful for semantic-enabled geospatial information retrieval: subsumption reasoning and similarity measurement. However, while the former one can be applied to query service ontologies described in OWL-S or WSMO/WSML, most existing similarity theories are not able to cope with logic-based service descriptions. This chapter presents initial results in developing a directed and context-aware similarity measure that compares WSML concept descriptions for overlap and therefore supports retrieval within the upcoming Geospatial Web.*

24.1 Introduction and Motivation

The idea of the Web service-oriented architecture (SOA) is based on the publish-find-bind pattern. To make a service available on the Internet, the provider has to publish relevant metadata to a service broker. Next, a requestor can discover (find) registered services and establish a connection (bind) to them. From a syntactical point of view, the SOA-Stack offers specifications for each part of the pattern: WSDL for Web service description, UDDI as a repository for description, discovery and integration and SOAP as protocol for service binding. However, to enable semi-automatic service discovery, i.e., to specify the capabilities of Web services and search queries in an unambiguous and computer-interpretable way, a semantic-enabled markup language becomes necessary. Moreover, beside this common language, a framework needs to be defined specifying which mandatory and optional metadata should be annotated. From the provider's perspective, service ontologies described using OWL-S[107] or WSMO (WSMO 2005a) satisfy these requirements. A detailed comparison is discussed in (WSMO 2005b); note, however, that it is written from the perspective of the WSMO community. Both approaches define functional and nonfunctional service properties, service grounding (binding) and a semantic-enabled annotation language. Although they specify what has to be said about a service, the definition of an adequate semantic search paradigm is out of their scope.

Over the last years of research, subsumption reasoning and similarity measurement turned out to be applicable for geospatial information retrieval. The idea behind subsumption-based retrieval (Lutz and Klien 2006) is to rearrange a queried application ontology taking a search concept into account and to return a new taxonomy in which all subconcepts of the injected search phrase satisfy the user's requirements. However, using this approach forces the user to ensure that the search concept is specified in a way that it is neither too generic (and therefore at a top level

of the new hierarchy) nor too specific to get a sufficient result set. In fact, the search concept is a formal description of the minimum characteristics all retrieved concepts need to share. Moreover, no measurement structure is provided to answer which of the returned concepts fits best. However, this is not necessarily a critical point, because all subconcepts at least share the requested properties. In contrast, similarity computes the degree of overlap between search and compared-to concepts and, as measurement structure, provides a (weak) order. Both characteristics turn out to be useful for information retrieval and matching scenarios. On the one hand, the determination of conceptual overlap simplifies phrasing an adequate search concept, and on the other hand, the results are ordered by their degree of similarity to the searched concept. Similarity-based retrieval does not necessarily imply a subsumption relation between search and compared-to concepts (see Figure 24.1); in some cases even disjoint concepts may be similar to each other (e.g., mother, father). In contrast to subsumption-based retrieval, the search phrase typed into the system is not an artificial construct, but the concept the user is really looking for in the external service ontology without presuming that all returned concepts share a specific property.

In other words, the benefits similarity offers during information retrieval, i.e., to deliver a flexible degree of conceptual overlap to a searched concept, stand against shortcomings during the usage of the retrieved information, namely that the results do not necessarily fit the user's requirements. To make the difference between both approaches more evident, one can imagine a search phrase specified by using a shared vocabulary (see Figure 24.1) to retrieve all concepts whose instances *overlap* with waterways. In contrast to the subsumption-based approach, similarity measurement would additionally deliver concepts whose instances are located *inside* and *adjacent* to waterways and indicate through a lesser degree of similarity that these concepts are close to, but not identical with, the user's intended concept.

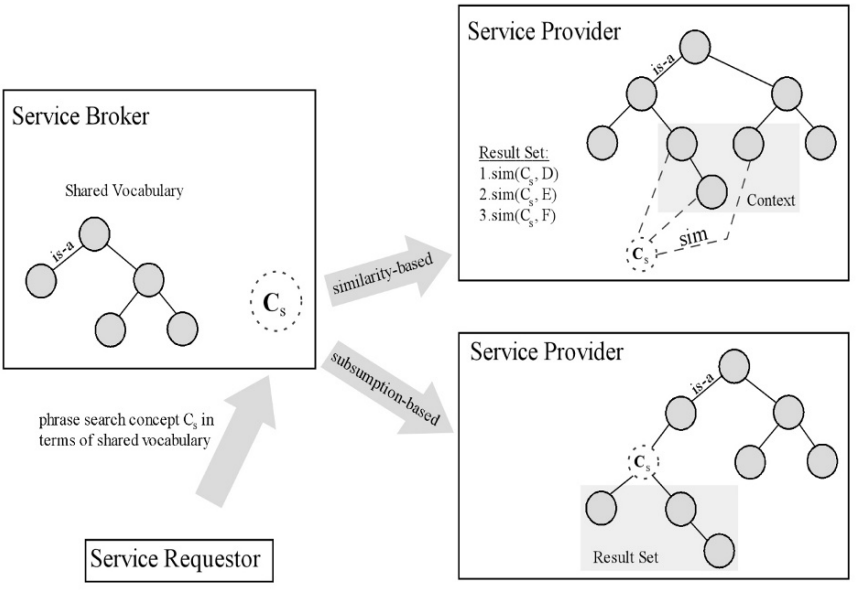


Figure 24.1: Subsumption and similarity-based retrieval using a shared vocabulary

Following the above argumentation, similarity supports users during information retrieval; however, this presumes that the chosen similarity measure supports the representation language of the inspected service (ontology). It turns out that, besides the fact that several similarity theories make fundamentally different assumptions about how and what is measured (e.g., feature vs. geometric model (Goldstone and Son 2005), most of them come with their own proprietary knowledge representation format. In contrast, the majority of service ontologies are specified using standardized or commonly agreed upon logic-based knowledge representation languages and especially various kinds of description logics. This leads to a gap between available similarity theories and existing ontologies that oppose a wider application of similarity measures as part of the Geospatial Web.

Additionally, most proprietary knowledge representation formats associated with existing similarity theories lack a formal semantics and also language constructs proven to be useful for conceptualization (such as relation-filler pairs). This is a crucial issue because in computer science the concepts between which similarity is measured are representations of the concepts in our minds. Consequently, the lack of a precise and expressive representation language impacts the quality of the resulting similarity assessments, as discussed in Janowicz (2005), for the feature-based MDSM theory (Rodriguez and Egenhofer 2004). The same arguments hold for geometric approaches to similarity based on Gärdenfors' idea of conceptual spaces (Gärdenfors 2000). To integrate relations and hence improve the expressivity of conceptual spaces for similarity measures, Schwering (2005), for instance, combines the geometric approach with classical network models. Initial approaches towards similarity measures for expressive description logics are discussed in d'Amato et al. (2005) and Janowicz (2006). A theory applying similarity for Web service comparison based on OWL-S is presented in Hau et al. (2005); however, it does not take neighborhood and alignment models into account. An overview about existing similarity theories, their application areas and characteristics is out of the scope of this chapter and was recently discussed in Goldstone (2004).

This chapter presents initial results on how similarity measurement can support semi-automatic information retrieval and matching tasks within the Geospatial Web.

24.2 Similarity between WSML Concept Descriptions

This section describes the proposed similarity measurement framework focusing especially on attribute-filler (respectively, relation-filler) similarity. Starting with a service integration scenario, the used representation language (WSML-Core) will be introduced and the similarity framework will be discussed step by step.

24.3 Scenario

A European lodging portal on the Internet is providing information about accommodations in cities attractive to tourists. To avoid maintenance costs the service provider does not store the information in a local database, but dynamically connects to external Web services. However, to offer a consistent interface and vocabulary to the portal users, the service provides its own terminology. To do so, the types of accommodations distinguished in the external services have to be aligned to the local terminology. One of the external services for delivering information about accommodations in Amsterdam provides separate conceptualizations for houseboats and botels (e.g., Amstel Botel Amsterdam),[108] while the local ontology does

not make this distinction. The task of similarity measurement within this scenario is to propose whether botels should be displayed as houseboats, hotels or youth hostels within the local terminology presented via a Web interface to the user. The provider therefore runs a similarity query against the local ontology using the external concept *botel* as search phrase (C_s). In addition, the service provider specifies a search context, i.e., a description of the minimum requirements all compared-to concepts need to fulfill (to be housings in this case). The result of the query is a list of similarity values indicating how close the compared conceptualizations are. It is assumed that both the external service and the accommodation portal stick to a shared vocabulary (Figure 24.1) that specifies the base concepts of the domain and that the concepts *botel*, *houseboat*, *hotel* and *youth_hostel* are defined in terms of this shared vocabulary.

24.4 WSMO and WSML

As similarity between concepts is based on their specification and the chosen representation language, this section gives a brief overview about the Web Service Modeling Language (WSML) and introduces simplified conceptualizations for the types of accommodations distinguished in the scenario.

Based on the Web Service Modeling Framework (WSMF) developed by Fensel and Bussler (Fensel 2002), the Web Service Modeling Ontology (WSMO) specifies four main modeling elements describing various aspects of Semantic Web services needed within the publish-find-bind pattern:

- **ontologies** providing the formal semantics for goals, Web services and mediators and linking human and machine terminology together,
- **goals** specifying user aims with respect to the requested service functionalities,
- **Web services** representing the offered functionality in terms of its capabilities and nonfunctional properties,
- **mediators** offering several methods to overcome interoperability problems.

While WSMO describes what needs to be said, WSML (WSML 2005) is the corresponding modeling language providing a formal syntax and semantics to describe these elements in a machine-interpretable and unambiguous way. It supports both a condensed machine oriented as well as a human-readable syntax and comes in five flavors of different expressivity: WSML-Core, WSML-DL, WSML-Flight, WSML-Rule and WSML-Full. Independent from a certain variant, WSML distinguishes between the following language elements: concepts (and their attributes), relations, instances (of concepts and relations) and axioms. However, the abilities to describe them depend on the chosen WMSL flavor. For each element, additional nonfunctional properties, mostly taken from the Dublin Core schema, can be specified.

The similarity measurement framework introduced in this chapter is defined for the WSML-Core variant based on the intersection of description logics with logic programming and acts as a base and exchange vocabulary for WSMO. In WSML-Core the usage of relations is restricted to binary predicates, and cardinality restrictions are not supported. The WSML documentation recommends using concept attributes instead of relations wherever possible. Moreover, WSML-Core does not allow for specifying the attribute features transitive, symmetric, reflexive and inverseOf within local concept descriptions. However, they can be added as global axioms to the service ontology and linked to the intended concept via the Dublin Core element dc:relation (WSML 2005, 27). Although WSML distinguishes between

constraining (ofType) and inferring (impliesType) attribute and relation descriptions, the former can only be applied to data types within the Core variant (WSML 2005, 17). WSML offers built-in data types, such as strings, integers, doubles or dates, which correspond to XML Schema data types and operators (XQuery functions) such as equal or numericGreaterThan. The syntax and semantics (mapped to Horn Logic) of the language constructs used within WSML-Core as well as an exemplary concept definition are depicted in Table 24.1 (see WSML 2005; see pp. 27–30 for further details).

Table 24.1: Syntax and semantics of WSML-Core

WSML-Core (syntax)	Horn Logic (semantics)	Example
(*head* **impliedBy** *body*.)	(*head*) $\Leftarrow$ (*body*)	Youth_Hostel **subConceptOf** {Building, Housing} **nonFunctionalProperties** dc#description **hasValue** 'concept of a youth hostel' category **ofType** _integer service **impliesType** SelfService offers **impliesType** Room
(*lexpr* **or** *rexpr*)	(*lexpr*) $\vee$ (*rexpr*)	
(*lexpr* **and** *rexpr*)	(*lexpr*) $\wedge$ (*rexpr*)	
(*X1* **memberOf** *id2*)	*id2*(*X1*)	
(*id1* **subConceptOf** *id2*)	*id2*(x) $\Leftarrow$ *id1*(x)	
(*X1*[*id2* **hasValue** *X2*])	*id2*(*X1*,*X2*)	
(*id1*[*id2* **impliesType** *id3*])	*id3*(y) $\Leftarrow$ *id1*(x) $\wedge$ *id2*(x,y)	
(*id1*[*id2* **ofType** *dt*])	*dt*(y) $\Leftarrow$ *id1*(x) $\wedge$ *id2*(x,y)	
($p(X_1,...,X_n)$)	$p(X_1,...,X_n)$	

Although we stick to the human-readable syntax within this chapter, it has to be mentioned that compared WSML concepts have to be preprocessed before similarity is measured. The necessary steps are described in WSML (2005, 42f) and result in a WSML normal form (see also Janowicz 2006). The underlying idea is to decompose complex descriptions to simple ones. Note that concepts inherit all attributes specified for their ancestors.

Table 24.2 shows possible conceptualizations for the types of accommodations described in the scenario. In contrast to Hotel and Youth_Hostel (see Table 24.1), Botel and Houseboat are defined as subconcepts of Boat; however, houseboats are usually self-serviced and are rented as a whole and not per room.

Table 24.2: Conceptualizations for the Botel-Houseboat scenario

Botel	Houseboat	Hotel
subConceptOf {Boat, Housing} category **ofType** _integer service **impliesType** Service offers **impliesType** Room borders(i) **impliesType** Waterway	**subConceptOf** {Boat, Housing} category **ofType** _string service **impliesType** Self-Service inside **impliesType** Waterway	**subConceptOf** {Building, Housing} category **ofType** _integer service **impliesType** Service offers **impliesType** Room

Note that *borders(i)* (borders from inside) corresponds to TPP and *inside* to NTTP in RCC8 (Cohn 1997); however, these relations need more investigation for 3D spatial neighborhoods (Kuhn 2002).

24.5 Similarity Measurement Framework

The presented theory measures similarity between concepts (in normal form) by stepwise comparing their WSML-Core descriptions, where a high level of overlap indicates high similarity and vice versa. To do so, all available language constructors, i.e., subConceptOf/subRelationOf and attribute (respectively, relation) as well as the restrictions for their fillers by ofType and impliesType, have to be taken into account. Similarity (sim) is therefore defined as a polymorphic, binary and real-valued function $\mathbf{X} \times \mathbf{Y} \rightarrow [0, 1]$ providing implementations for all language constructs. The overall similarity (sim_o) between concepts is just the normalized (and weighted) sum of the single similarities calculated for all compared-to parts of the concept descriptions. A similarity value of 1 indicates that compared concepts are equal, whereas 0 implies total dissimilarity. In the following σ denotes the normalization factor while ω is used to represent weightings.

In general, a similarity measurement framework consists of the following five phases – their concrete implementation and relative importance, however, depend on the chosen representation language: (i) define search concept and context; (ii) generate canonical normal form for compared concepts; (iii) align parts for comparison; (iv) apply similarity functions to compared-to parts; (v) derive normalized overall similarity.

Preprocessing steps to derive a WSML-Core normal form (phase 2) have been discussed in Section 24.4 and are therefore not considered here in further detail. A more complex example pointing out the importance of canonical representation for similarity measurement is discussed in Janowicz (2006).

24.5.1 Search Concept and Context

As depicted in Figure 24.1, a search concept (also called a source) is phrased in terms of a shared vocabulary and compared to the concepts (called targets) in an examined ontology. The search concept needs to be specified in the same representation language as the target concepts, or mappings between the languages have to be defined. The target concepts are not necessarily just all concepts in the examined ontology but defined by a search context. The idea of context [see also the Matching Distance Similarity Measure MDSM (Rodriguez and Egenhofer 2004)] is on the one hand to determine which parts from the service ontology have to be compared to the search concept and on the other hand to influence the measured similarity, making it situation-aware. Within the presented approach context is used to combine the benefits of subsumption reasoning and similarity-based retrieval. It is defined as a set of concepts from the examined application ontology, which are are subconcepts of $C_{lcs:}$ context = $\{C| \; C \sqsubseteq C_{lcs}\}$ after reclassification (Lutz and Wolter 2006). In other words, context determines the universe of discourse [called "application domain" in Rodriguez and Egenhofer (2004)]. In the presented accommodation scenario, C_{lcs} guarantees that all concepts proposed to be similar to Botel at least act as accommodations (subconcepts of Housing). Therefore, similarity to cargo ships or ferries would not be measured, although they are kinds of boats as well.

24.5.2 Alignment Matrix

After their expansion to WSML-Core normal form, concepts are lists of attributes, respectively, relations (with restrictions for their fillers), including those inherited from their ancestors. While search concept and context define concepts to compare, it has to be determined which parts (e.g., which attribute-filler pairs) of the selected concepts are compared to each other. To do so, a matrix $C_s \times C_t$ of all possible combinations is generated. Similarity can only be computed between the same kinds of language elements, i.e., attribute-filler pairs using the ofType keyword are not compared to those using impliesType and so on. Therefore, such pairs are not further taken into account. Next, the following steps are applied for all parts of the source concept description and each part of C_s and C_t is only selected once:

- If the matrix contains an identical attribute/relation-filler pair for the search and the target concept, the similarity for this pair is 1 and the normalization factor σ is increased by 1.
- If the matrix contains an attribute/relation-filler pair out of the target concept description where the attribute/relation is identical to the pair in the source concept but the fillers are different, similarity between the fillers is calculated. If there are more such pairs, the one with the highest similarity for the filler is selected and σ is increased by 1.
- If for an attribute/relation-filler pair out of the source concept description no pair with an identical attribute/relation could be found, the most similar pair is selected, where the similarity between the attributes/relations can be determined using a conceptual neighborhood graph; σ is increased by 1.
- If no neighborhood graph is specified for the compared-to attributes/relations, their similarity is measured by co-occurrence and σ is increased by 1.
- Co-occurrence is also determined for superconcepts if they are base symbols (primitives) defined in the shared vocabulary and therefore have no description to be compared (inherited); σ is increased by 1.
- For parts of the search concept that could not be compared, similarity is 0 and σ is increased by 1; while σ is not increased for parts of the target concept that could not be aligned to corresponding parts of the search concept.

In other words, for each part of the search concept's description, a counterpart from the compared-to concept's description is chosen in a way that a meaningful similarity can be computed between them afterwards and each part is only examined once. The alignment phase is directed, i.e., asymmetric (Rodriguez and Egenhofer 2004), in a sense that the resulting overall similarity depends on the search direction. Therefore, $sim_o(C_s, C_t)$ is not necessarily equal to $sim_o(C_t,C_s)$. While each element of the search concept's description is compared to an element from the compared-to concept, some parts of the latter may remain uncompared. This is always the case if the target concept is specified by more elements than the search concept. The similarity value for these remaining parts is always 0 while they do not increase the normalization factor σ. If, however, the search concept is described by more elements than can be compared, σ is increased by 1 for each remaining part. This decreases the overall similarity. If the target concept in the examined ontology is more specific than requested by the user (via the search concept), this has no impact on the measured overall similarity. On the other side, similarity decreases if the user's search concept is more specific than its counterpart in the examined ontology.

24.5.3 Similarity Functions

After determining which parts of the search and target concept are compared to each other, the similarity between these selected parts is calculated. Two situations have to be taken into account here: the similarity between attribute/relation-filler pairs depends on both the similarity between the attributes (respectively relations) and the similarity between the fillers. If the fillers are *concepts*, the similarity framework is recursively applied to the compared fillers, i.e., an alignment matrix is created for the compared concepts and similarity for the selected parts is measured. If, however, the fillers are *data types* (expressed via ofType keyword in WSML-Core), similarity is determined via a matching function and no recursion is necessary.

$$sim_{cf}(ac_s, ac_t) = sim_a(a_s, a_t) * sim_o(c_s, c_t) \quad \textbf{(1)}$$

Equation (1) shows how the attribute-filler similarity (sim_{cf}) is calculated for concept fillers, where sim_a is the similarity between attributes and sim_o is the overall similarity between their fillers (similarity for relations-filler pairs is calculated accordingly, but is omitted here for reasons of readability). In addition to this multiplicative approach, similarity could also be defined as the (weighted) average between attribute and filler similarity, which is discussed in Section 24.3.

According to the alignment matrix, similarity between attributes (sim_a) can be determined in two ways: using a conceptual neighborhood graph (ndw), as depicted in Figure 24.2 for a topological neighborhood, or via co-occurrence (sim_{co}). The benefit of conceptual neighborhoods is that they imply a very natural notion of similarity – which is just the inverse and normalized graph distance between the compared attributes [see Eq. (2)]. Although the edge weightings may vary with respect to the chosen conceptual neighborhood (n) or intersection matrix, they are usually set to 1 per edge and symmetric [see also (Bruns 1996; Li 2006) for similarity measures between spatial scenes].

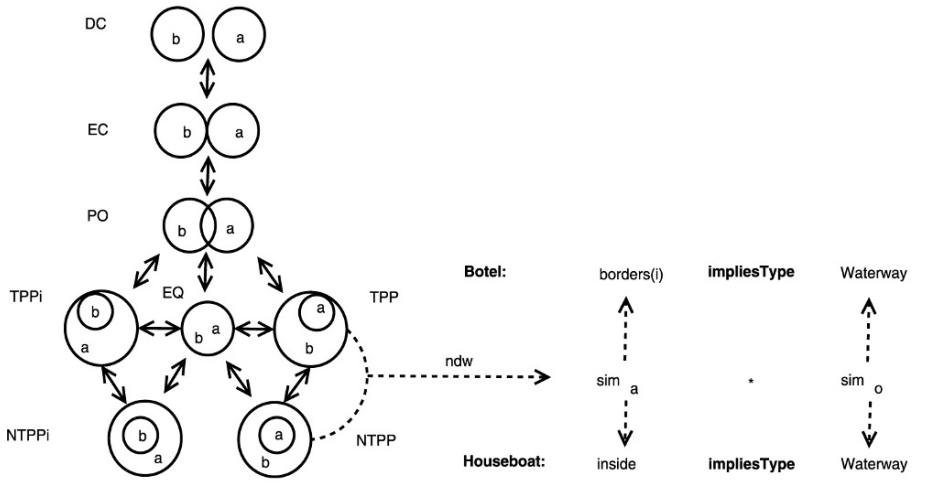


Figure 24.2: Spatial neighborhood distance (Cohn 1997) and interattribute similarity

Equation (2) describes how similarity between attributes is determined via their relative distance within a conceptual neighborhood, where $distance_n$ between a_s and a_t is the shortest path through the graph while $max\_distance_n$ represent the longest path.

$$ndw(a_s, a_t) = \frac{max\_distance_n - distance_n(a_s, a_t)}{max\_distance_n} \quad \textbf{(2)}$$

In contrast, the co-occurrence (also called common subsumee approach) assumes that attributes are more similar if they share more common subattributes. In Eq. (3), sim_{co} is defined as the ratio between the number of subsumees of both attributes and the number of subattributes of one or both of them (and all y are elements of the context). This notion of co-occurrence is comparable to the Jaccard similarity coefficient (Tan et al. 2005). Note that within WSML-Core the subattribute relationship is specified as implication using logical expressions (WSML 2005, 29). The letters x and y are chosen here to indicate that the same equation is applied to attributes, relations and primitive concepts (base symbols) as well.

$$sim_{co}(x_s, x_t) = \frac{|\{y | (y | x_s) * (y | x_t)\}|}{|\{y | (y | x_s) + (y | x_t)\}|} \quad \textbf{(3)}$$

As can be seen from Eq. (1), the attribute-filler similarity sim_{cf} calls the overall similarity sim_o to determine the overlap between involved concept fillers. The overall similarity, however, again invokes sim_a to compare the attributes specified for these concepts and so on. The process terminates when the concepts specified as fillers have no concept description, i.e., are base symbols (primitives) of the shared vocabulary. According to Section 24.2.3.2, their similarity is determined via Eq. (3), where the subsumees are not attributes but subconcepts. The same approach is applied if superconcepts defined in the head of concept definitions are base symbols and therefore do not bequeath attributes to their subconcepts.

While the former paragraphs focused on concepts as fillers, the similarity (sim_{df}) between attribute-filler pairs with data-type fillers is determined according to Eq. (4). The function match() returns 1 for the same type or if all instances of d_s could be converted to d_t without losing information (respectively, precision; such as from integers to decimals) (WSML 2005; p.88); otherwise match() returns 0. Some problems related to similarity with respect to data types are discussed in the section on further work.

$$sim_{df}(ad_s, ad_t) = sim_a(a_s, a_t) * match(d_s, d_t) \quad \textbf{(4)}$$

24.5.4 Overall Similarity

Finally, the overall similarity (sim_o) between search and target concepts is the normalized sum of the similarities derived by comparing attributes with concept fillers (via sim_{cf}), attributes with data-type fillers (via sim_{df}) and primitive concepts (base symbols) in the head of c_s and c_t (via sim_{co}). In Eq. (5), (c_{si}, c_{ij}) represent the parts of the source and target concepts selected for comparison within the alignment matrix AM_{si}.

$$sim_o(c_s, c_t) = \frac{1}{\sigma} \sum_{(c_{s_i}, c_{t_j}) \in AM_{st}} sim(c_{s_i}, c_{t_j}) \quad (5)$$

24.6 Human Subject Testing

This section describes the results from a Web-based human subject test, developed to examine how users rate the similarity between attribute/relation-filler pairs. After explaining the goals of the test, subjects were asked to make similarity estimations using a slider that ranges from very dissimilar to very similar (which corresponds to a value range between 0 and 100). The slider was situated between the compared entities, and its start position was halfway between both. The test consists of three steps, each containing four pairs to be compared. In the first step, subjects were asked to rate similarity between spatial relations such as *disjoint* and *meets*. In the second step, object pairs such as *waterway* and *river* were compared. Finally, in the third step, subjects had to rate the similarity between combinations of both (e.g., *disjoint waterway – meets river*). These similarity estimations were than compared to automatically generated similarity values using three different approaches: the average, a weighted average with flexible weightings and the multiplicative approach depicted in Eq. (1). The necessary attribute and filler similarities were taken from the first two steps of the test.

Out of 84 similarity estimations derived from step 3, 80 were taken for further computation. As depicted in Figure 24.3, the multiplicative approach produces the best results. In 41 of the 80 cases, the absolute deviation did not exceed 10 points; however, the approach tends to underestimate in general. In contrast, the weighted average tends to overestimate and the results are not as precise (33 of 81). The simple average approach was always overestimating and the deviation from human's estimations was high in general.

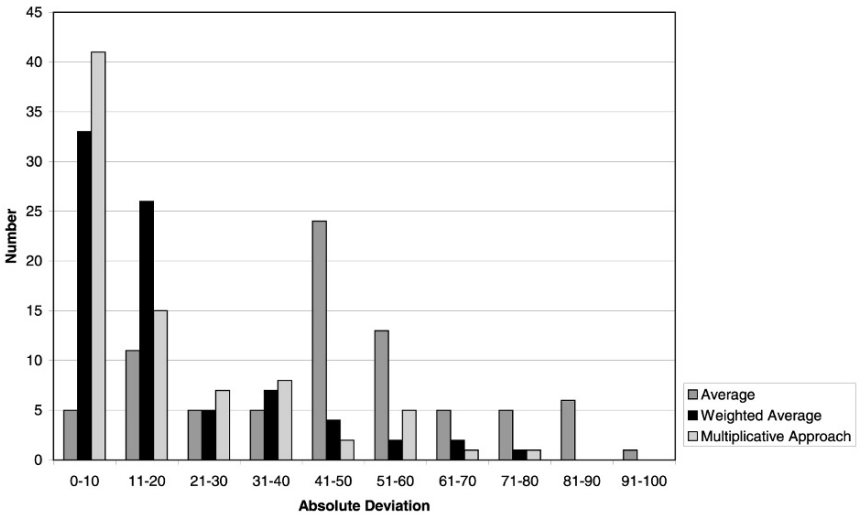


Figure 24.3: Absolute deviation between machine and human similarity estimations

To explain the idea of similarity estimations to the subjects, they were told that comparing relation-object pairs could be imagined as rating how probable two people (describing a certain situation in different words) actually talk about the same situation or not. It turns out that this explanation may be a reason why some of the human's similarity estimations were inconsistent and neither captured by the multiplicative approach nor the weighted average. While *inside-disjoint* and *lake-channel* were rated to be dissimilar, the combination was rated to be more similar than expected. Subjects assumed that if a described object is inside a lake, it is disjoint from a channel.

24.7 Discussion and Further Work

The directed and context-aware similarity theory presented within this chapter is able to measure the overlap between concepts specified using WSML-Core and can therefore support integration and retrieval within service-oriented architectures. In contrast to previous work, it points out possible ways of combining subsumption reasoning and similarity. Nevertheless, a lot of work remains to be done to apply these initial results to sophisticated real-world applications.

Referring to the accommodation scenario, it turns out that botels are more similar to hotels (0.67) than to houseboats (0.62) or youth hostels (0.5). However, the measured similarities depend on the representation of the compared concepts within the provider's ontology. Services focusing on vessels instead of accommodations may use different conceptualizations, making botel and houseboat more similar. Note that from now on the accommodation service can also display botels on the portal's Web site whenever a user is looking for hotels in Amsterdam, but (in contrast to subsumption-based retrieval) integrating the concept botel into the local knowledge base would lead to inconsistencies (a botel is not a building).

It turns out that while the comparison of attributes (respectively, relations) restricted by concept fillers is well examined (d'Amato 2005; Schwering 2005; Janowicz 2006), the question of how to develop a meaningful theory for data-type similarity still remains unsolved. One of the main reasons is missing information about the level of measurement or nonlinear measures (Schade 2005). For instance, the category of a hotel is measured in stars and represented as an integer on an ordinal scale, while the distance to a beach is also of the data-type integer but on an interval scale: 100 m to the beach is half as much as 200 m, but a 2-star hotel is not half as good as a 4-star hotel. In addition, according to Eq. (4), the match function returns 0 for comparing decimals to integers, although the lost precision may not be relevant for a user in a certain situation. Taking complex XSD types into account would further complicate the determination of a meaningful notion of data-type similarity (e.g., xs:sequence).

Another important issue is the extension of the presented approach to cope with more expressive WSML variants. The major question arising here is what can be said (in terms of similarity) about compared logical expressions (e.g., via generalization). While the presented theory demonstrates how to compare concepts within WSML service ontologies, mediators, goals and capabilities were not discussed within this approach. However, further theories may benefit from the idea of WSML mediators as mapping rules (WSMO 2005a). Moreover, it has to be examined how users, such as the service provider, can phrase search concepts without being domain experts and trained logicians. Finally, further refined human subject tests are necessary.

Chapter 25

Geospatial Data Integration with Semantic Web Services: The eMerges Approach

Vlad Tanasescu • Alessio Gugliotta • John Domingue • Leticia Gutiérrez Villarías • Rob Davies • Mary Rowlatt • Marc Richardson • Sandra Stinčić

Abstract. *Geographic space still lacks the semantics allowing a unified view of spatial data. Indeed, as a unique but all encompassing domain, it presents specificities that geospatial applications are still unable to handle. Moreover, to be useful, new spatial applications need to match human cognitive abilities of spatial representation and reasoning. In this context, eMerges, an approach to geospatial data integration based on Semantic Web Services (SWS), allows the unified representation and manipulation of heterogeneous spatial data sources. eMerges provides this integration by mediating legacy spatial data sources to high-level spatial ontologies through SWS and by presenting for each object context dependent affordances. This generic approach is applied here in the context of an emergency management use case developed in collaboration with emergency planners of public agencies.*

25.1 Introduction

Web 2.0 applications, by offering large amounts of resources to users for small fees, by weaving social networks where only forests of text-based hyperlinks existed and by providing desktop-like applications to the browser, are changing the way we interact on the Web. Part of this evolution is a renewal of the available mapping applications; closed, static and symbolic traditional Web map applications are progressively being replaced by *Web 2.0 maps* employing new means to achieve a *map reality effect*, which is the ongoing effort of rooting the maps into the cognitive reality by giving more natural-looking insights into the geography covered by it. Also, by freely distributing APIs, new Web 2.0 maps lead to an explosion of mash-ups, minimal applications developed by independent technically skilled users that aggregate data in a spatial context in order to fulfill a specific goal.

The popularity of Web 2.0 maps and mash-up applications[109] shows the interest and the appeal of the geographic environment for Web users; mash-ups are used for such a wide variety of goals that it seems that space, mediated through realistic Web maps, may provide the terrain for data integration rooted into human cognition that the more abstract textual Web has not yet succeeded to achieve.

However, mash-ups, as isolated attempts at data integration, do not have to cope with the semantic complexity of multiple heterogeneous data sources; usually, the service providing the data is integrated by the developer as a single and isolated map layer, making the related semantics clear.

To allow large-scale integration, semantic descriptions are needed (Egenhofer 2002; Kuhn 2005). Semantic Web Services (SWS) are the result of an acknowledgement that Web Service technology (WS), even in its standardized form, cannot achieve a satisfying level of interoperability without appropriate high-level seman-

tics. Indeed, WS based on ad hoc REST APIs or on standards such as UDDI[110] for discovery, WSDL[111] for interface description and SOAP[112] for message passing, simplify the task of the developer but without dismissing his or her knowledgeable intervention. Particularly when new services are to be integrated to an application, developers need to study the WS descriptions to match inputs, outputs and invocation workflows with the existing systems. By using SWS, if the vision of fully automatic interaction and composition is still a research question, the following tasks are already greatly alleviated:

- *Discovery* of useful services is achieved by matching a formal task description against SWS' semantic descriptions.
- *Mediation* between heterogeneous services can be specified at the level of data format, message protocol and business processes.
- *Composition* of services provides a means of creating a new service by aggregating existing components.

IRS-III (Cabral et al. 2006), a platform and broker for developing and executing Semantic Web Services, adopts a Semantic Web approach based on ontological descriptions, expressed formally in OCML (Motta 1999). In particular, IRS-III incorporates and extends the Web Services Modeling Ontology (WSMO) (Roman et al. 2005). *Goals*, a concept existing in WSMO to describe users' needs as distinct from specific WS functionalities, can be invoked in this extension, which ensures a more intuitive way of interacting with clients in a Semantic Web (SW) context.

The eMerges approach applies SWS technologies to the Geospatial Web, which has been designed as an e-government use-case domain in the context of the DIP project (funded under the European Union's IST Program FP6). eMerges illustrates the way in which spatially related data delivered through SWS can ease the management of specific use cases by aggregating data originating from different sources and presenting them in a way that is both consistent and task relevant.

We first describe how SWS applications are built using IRS-III, giving an example of how to use SWS, then briefly present the specificities of Geographic Space as well as the eMerges generic approach to handling spatial objects in context, and finally, before concluding, discuss functionalities of the eMerges prototype implementation.

25.2 IRS-III and SWS Applications

Applications using IRS-III follow a layered approach (see Figure 25.1) in which (micro-) functionalities of legacy systems are exposed through Web Services – based on standards or on REST – and described with ontologies. These Semantic Web Services can then be invoked from the (Web) presentation layer, by using a provided API, SOAP messages or the REST protocol.

The Web Service Modeling Ontology (WSMO) (Roman et al. 2005) is a formal ontology for describing the various aspects of services to enable the automation of WS discovery, composition, mediation and invocation. The meta-model of WSMO defines four top-level elements: *ontologies*, *goals*, *Web Services*, and *mediators*. *Ontologies* (Gruber 1993) provide the foundation for describing domains semantically. They are used by the three other WSMO components. *Goals* define the tasks that a service requester expects WSS to fulfill. In this sense they tend to reflect the service user's intent.

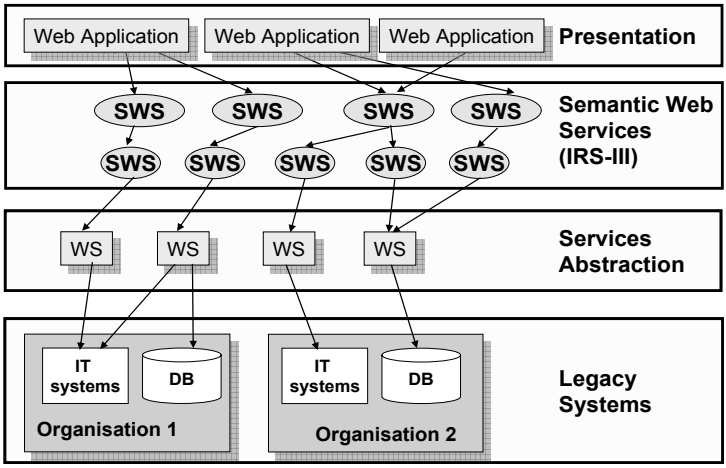


Figure 25.1: Generic architecture used when creating IRS-III-based applications

Web Service descriptions represent, in terms of *capabilities* (what the service can do) and *interface* (how to use it), the behavior of a deployed Web Service. The description also indicates how WS communicate (*choreography*) and how they are composed (*orchestration*). *Mediators* handle issues of data and process interoperability that arise between heterogeneous systems. One of the characterizing features of WSMO is that all components – ontologies, goals and Web Services – are linked by mediators. In particular, WSMO provides four kinds of mediators:

- *oo-mediators* for mediating between heterogeneous ontologies;
- *ww-mediators* connect WS to WS;
- *wg-mediators* connect WS with goals;
- *gg-mediators* link different goals, solve input conflicts and transform processes.

By extending WSMO's goal and Web Service concepts, clients of IRS-III can invoke Web services via goals. That is, IRS-III supports *capability-driven* or *goal-driven* service invocation, which allows the user to use only generic inputs, hiding the possible complexity of a chain of heterogeneous WS invocations. The decoupling of the actual user vision of a task and its execution allow us to get closer to the user's cognition of the situation and task. Mediators link goal and Web Services, solving existing mismatches and allowing complex composition of services.

The implementation use case was designed with the *Essex County Council (ECC)*. The ECC is a large local authority in southeast England. Following several interviews with spatial data holders in the ECC, it was decided to focus the scenario on the ECC Emergency Planning Department and, precisely, on a previous emergency situation: the snowstorm that occurred in the vicinity of Stansted Airport on January 31, 2003. Because of the snow, drivers were trapped for several hours in their cars on the M11, a motorway in the UK; as a result, access to Stansted Airport was difficult, and individuals required transport to nearby shelters and hospitals.

eMerges was used as the underlying conceptual framework to implement a decision support system assisting the Emergency Officer in handling the dynamics of the emergency situation and gather information related to a certain type of event, faster and with increased precision.

Data were integrated from three different sources. The UK's *Meteorological Office* provided snow-level information; *ViewEssex*, a centralized database maintained by British Telecommunications (BT) managed spatialdata for the ECC; and *BuddySpace*, an instant messaging client built on top of the *Jabber*[113] protocol, provided lightweight communication and collaboration means (Eisenstadt et al. 2003). Services were described by using domain ontologies that were mapped to integration ontologies. This process involved building goals and mediators to provide added value to the services, for example, through composition.

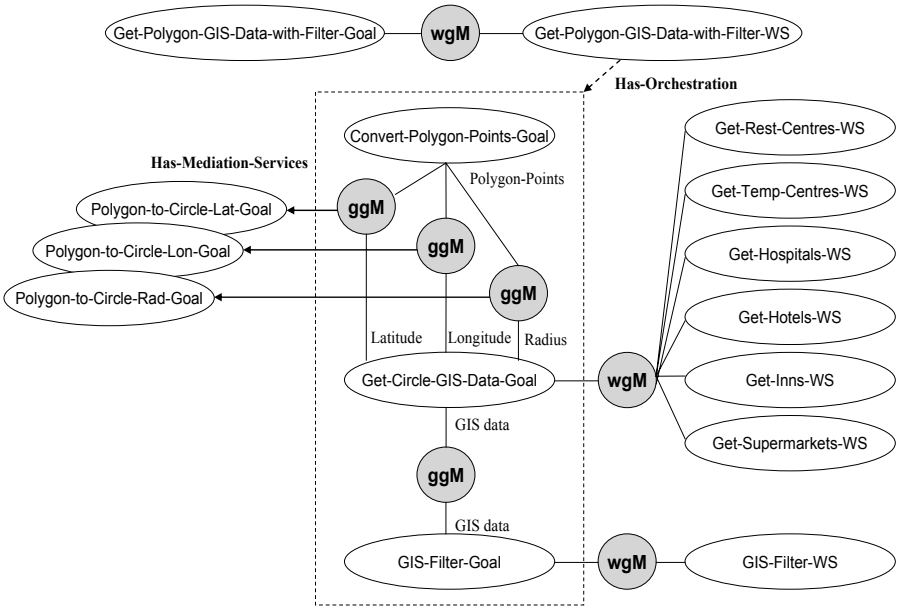


Figure 25.2: Structure of the WSMO description of the eMerges prototype; to avoid cluttering the diagram, wgM and Web Services balloons were omitted

To illustrate such a composition, we describe in the following the structure of the WSMO descriptions associated with an example goal, *Get-Polygon-GIS-data-with-Filter-Goal* (see Figure 25.2). This goal describes the request of a class of shelter (hospital, inn, hotel, etc.) in a delimited query area. The user selects a class of shelter, while the *polygon* query area is provided by the context. However, the only WS available returns a specific class of shelter in a *circular* query area. Moreover, results also have to be filtered in order to return only shelters relevant to the task (in our case, the management of a snowstorm emergency). Therefore, problems for invocation are (1) *selection* of the adequate WS, (2) *mediation* of the different area representations (polygon vs. circular), (3) *orchestration* of the retrieve and filter data operations. IRS-III offers different approaches to deal with these issues:

- *WS selection:* each WSMO description of WS defines, in its *capability*, the specific class of shelter that the service provides. All descriptions are linked to *Get-Circle-GIS-Data-Goal* by means of a unique *wg-mediator (wgM)*. The goal expects as input a class of shelter and a circular query area. At invocation time, IRS-III discovers through the wgM the WS associated to it. Then it selects one among them according to the specific class of shelter described in WS capabilities.

- *Area mediation and orchestration*: *Get-Polygon-GIS-data-with-Filter-Goal* is associated to a unique Web service that orchestrates – here, invokes in sequence – three subgoals. The first one simply gets the list of polygon edges from the input; the second is the above-mentioned *Get-Circle-GIS-Data-Goal*; and finally, the third invokes the smart service that filters the list of GIS data. The first two subgoals are linked by means of three *gg-mediators (ggM)* that convert the list of polygon edges provided by the first subgoal to the center (latitude and longitude) and radius of the circle that circumscribes that polygon. To accomplish this, we created three mediation services invoked through *Polygon-to-Circle-Lat-Goal*, *Polygon-to-Circle-Lon-Goal* and *Polygon-to-Circle-Rad-Goal*. The results of the mediation services and the class of shelter are the inputs of the second subgoal. A unique ggM connects the output of the second to the input of the third subgoal. No mediation service is necessary here.

Other improvements upon WS are made possible by IRS, such as describing complex orchestrations through a full work-flow model expressed in OCML; supporting data-flow and solving mismatches through mediators; and defining how to interact with a single deployed WS (e.g., policies) on the basis of a set of forward-chaining rules (Cabral et al. 2006).

25.3 The eMerges Approach

25.3.1 Semantics for the Geographic Space

It is well acknowledged that the spatial domain is somehow *special* (Peuquet 2002). Indeed, Geographic Space encompasses objects quite different from the ones we usually manipulate or are used to describing in knowledge bases; here scale, orientation, boundaries and cultural conceptions, among other elements, seem to matter to a greater extent (Smith and Mark 1999).

If a full review of the specificity of the geographic domain is beyond the scope of our work, three aspects of this specificity particularly oriented our research:

- *object/field divide*: it has been recognized that objects and fields – the assignment of values to spatial locations – have to co-exist in geographic applications (Couclelis 1992). However, this distinction still constitutes a problem for the object representation tradition. Indeed, *why* is an object such as a mountain a field or an object, or, better, *when* do we want it to be a field or an object? What about fields composed of other fields (e.g., land coverage)? If answering these questions in a generic manner is hard, human cognition never fails in choosing the best representation, object or field or composition of both, according to a context.
- *cognitive imperative*: space is experienced before being undestood, as shown by Naïve Geography (Egenhofer and Mark 1995), which demonstrates to what extent useful representations of space are to be rooted into human cognition. This is highlighted in yet another way by Web 2.0 maps, in which multiple reality effects are embedded, such as seamless continuity in map browsing instead of image by image retrieval, satellite imagery, road level or oblique photography, 2.5 or even 3D features. These representations are appealing since they allow the transition from the world of symbolic representation towards iconic models of reality used commonly in daily life, and therefore allowing applying cognitive models. These glimpses of a world behind the map provide us with new *affordances* (Gibson 1986) – what an element of the external world *allows me to do* as more essen-

tial than its other characteristics; symbols direct and focus the perception of affordances while the vision of realistic images allows the full range of them, *image schemata* (Mark 1989) – an element can be further reduced to simple concepts that are self-understandable; we are used to such kind of abstractions from perceptive inputs, or *conceptual spaces* (Gärdenfors 2000) – a concept as a point in a multidimensional space of simpler representations; the meaning of a real object is somehow defined by one's perspective on it.

- *multi-representation*: at the intersection of the object/field divide problem, and the need of cognitive approach to object representation, spatial applications need to represent spatial objects, objects that representation simply *changes* not only according to the level of detail needed or requested (*generalization*), but also depending on the task at hand. For example, an airport such as Stansted will be a node in a flight's graph from an international point of view, then become an independent region in a land cover study, or a simple traffic node, or a complex environment itself containing a road network and buildings, or a group of 3D structures with emergency access path in a fire escape scenario, etc. The multitude of contexts and corresponding relevant representations raises the question of the possible uniqueness of geographic object representation; indeed, if many representations are useful, how can they be linked and accessed in a timely manner, according to contextual information?

eMerges is an ongoing effort to address these concerns, by linking them via the notion of context. Indeed, in order to ultimately (i) alternate object and field representations, (ii) provide cognitively relevant information and (iii) choose between multiple representations of the same element, the representation of *spatial objects* becomes *context dependent*. We are going to define both notions in turn.

25.3.2 Spatial Objects

First, in order to describe and to reason about Spatial Objects in all their generality, a simple yet precise definition is needed. Our model is based on Galton's theory of objects and fields (Galton 2001), which defines a spatial object as belonging to a given *type* and having a *location* component (some "whereness") as well as *attributes* (also called *features*).

Mapping of arbitrary domain entities to spatial objects can be automatic or manual. In automatic mapping, a procedure collects each object's attribute value and transforms it into an attribute name/value pair of a spatial object, with a special treatment for id and location. In manual mapping, arbitrary transformations are possible. Once spatial objects are gathered, further mappings are needed to achieve independence between objects and their actual use.

For example, generalization is achieved in eMerges by using an *archetypes* ontology providing generic abstractions (e.g., *container*, *house*, *agent,* etc.) to which entities have to be mapped. In this way, even if the client application does not understand the type of element that is to be represented, a choice of representations and affordances is still possible by reasoning on the attached archetypes, which clients are requested to be aware of. A *hospital,* for example, can be represented as a *house*, the attached archetype, with affordances including how to get there, which is in any case sounder than other archetypal representations such as *agent* or *link*, which are distinct archetypes.

Moreover, to adapt the representation of a spatial object to a particular interface, the *HCI* ontology maps an object to a particular HCI representation. For example, some interfaces need "pretty names" selecting a feature to privileged display (e.g., on hovering on the object); an attribute of an adapted HCI concept allows us to specify which information, by automatic mapping (e.g., a procedure choosing any slot containing the string "id" or "name") or with a manual one.

These ontologies, together with the attached mapping mechanisms, are called *integration ontologies* since they allow the integration of spatially related data sources ranging over very different domains. Alone, they allow the integration of spatial data sources in a generic way; however, as the number of data sources increases, the task of presenting objects and possible queries according to the context, in order for the user not to be overwhelmed by the amount of information, becomes essential. Hence, the notion of spatial context becomes important in order to provide only relevant information and services.

25.3.3 Spatial Context

In order to alternate cognitively sound representations and actions, it is acknowledged that some extent of *context-awareness* (Dey and Abowd 2000) is needed. In eMerges, the main components of context are related to *user role*, *task*, *location* and *focus* of interest. Indeed, a user identifies herself as having a particular role, such as firemen responsible for transportation in a snowstorm emergency, or police forces responsible for a victim's accommodation. Moreover, weather information is available only in the region covered by the service, and the option of asking for it must be presented only in relation to objects related to weather investigation or emergency planning.

Object representations differ according to the context; e.g., emergency planners view shelters as points independently of scale, while the fire brigade responsible for transport needs precise access plans at a greater proximity. Second, to spatial objects are linked possibilities of action that allow getting more information in a precise context. For example, an area defined as an evacuation zone may offer goals allowing finding the nearest supermarkets – providing food – or hotels – providing accommodation – etc. This links the SWS notion of *goal* to the cognitive notion of *affordances* attached to an object. Therefore, when involved in a context, a spatial object receives specific affordances, linked to WSMO goals. Affordances allow navigation through the Geographic Space by successive and uniform information retrieval steps, i.e., as hyperlinks allow us to navigate from Web pages to Web pages, affordances are attached to an object depending on the context and allow retrieving additional spatial objects. For example, in a given context, a town object will afford retrieving nearby hospitals, while in another they will allow retrieving its administrative subdivisions. All retrieved objects are also captured within a similar context and present relevant affordances.

To achieve this, the question of whether a specific context reasoning engine has to be used is open. However, we believe that in the context of SWS, a more scalable solution may be achieved by distributing the task of context handling among *smart services,* which also implement reasoning in our architecture. Indeed, context pervades the elements of an SWS application and can be represented (i) at an affordances level, i.e., by offering very specific goals only, according to the context, e.g., a *get-heated-shelters* affordance will be presented in an emergency case involving low temperatures, or (ii) at a composition level, i.e., generic affordances are presented, but smart composition between goals ensures context relevance, e.g., the generic

affordance *get-shelters* is presented to the user but will highlight heated shelters according to the snowstorm task. The first solution has the advantage of being more explicit, whilst the second is easier to implement since it requires fewer goal definitions. Being able to handle context at every level makes both solutions possible in SWS-based applications.

25.4 Interaction in eMerges

The prototype implementation is a Web interface using Google Maps for the spatial representation part of the application. The interface is built using the Google Web Toolkit,[114] using AJAX techniques on the client to communicate with a Java servlet, which itself connects to IRS-III through its Java API. The most significant component of the interface is a central *map*, supporting *spatial objects*. A spatial object can have an area-based location, in which case it is displayed as a polygon, or a point-based one, in which case it is displayed as a symbol. All objects present the same interface, with *affordances* and *features*, displayed in a pop-up window or in a hovering transparent region above it (see Figure 25.3).

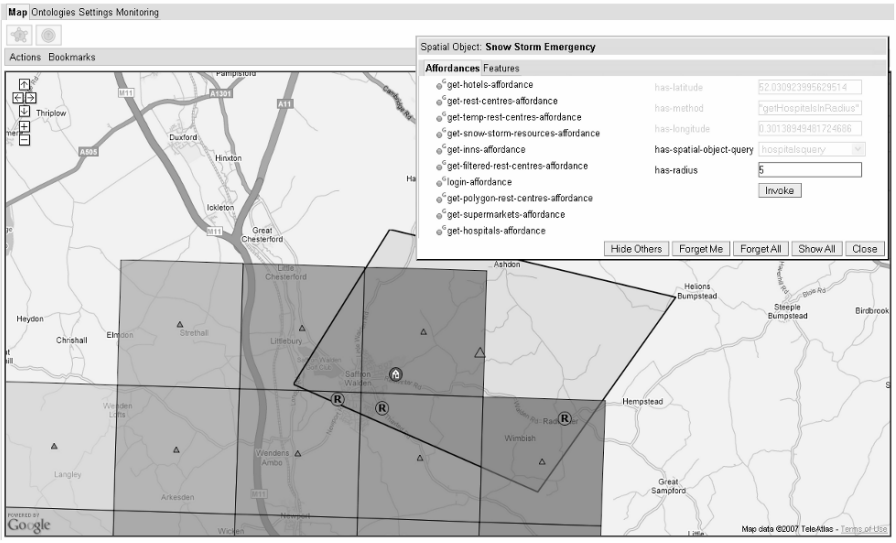


Figure 25.3: Screenshot of the eMerges implementation; a field (bottom left), point elements and a polygon (middle), affordances and features dialog (top right)

As an example of practical usage, we describe how an emergency officer (EO) gathers information regarding a possible emergency situation, and the affordances which are made available depending on the context. The procedure is as follows:

- Based on external information about the possibility of a weather emergency, the EO defines an area of interest on the map.
- A pop-up window containing a tree view appears showing elements of the available ontologies, of which location is an area, and which are therefore candidates to be the type of the region. The choice in this example is weather-investigation-area.

- As defined in the ontology, the new instance has attached *features* and *affordances*. One affordance allows users to log in into BuddySpace to ask field agents for more information, while another one accesses meteorological information.
- The EO requests snow information in the area.
- The result is a field of *snow-value* objects attached to regions. Although essentially representing scalar values, each polygon constitutive of the field is itself an object, presenting affordances and features.
- Depending of the results, the original investigation area becomes a *low-snow-hazard-emergency* or a *high-snow-hazard-emergency*, as described in the ontology.
- A high-snow-hazard-emergency provides affordances that allow the user to ask for more information, for example, to request all rest centers inside the region.
- Rest centers are retrieved with features and affordances.
- And can be used to get more context-relevant information, i.e., other resources nearby such as hospitals.
- The EO can also choose to log in to BuddySpace to contact the relevant persons to request action or information.

A screencast of similar interactions as well as a live version are available online,[115] to be used preferably with the Firefox Web browser.[116]

25.5 Discussion

Two main aspects of eMerges can be related to other approaches: data integration and context-based navigation of data.

Integration of new data sources is relatively simple in eMerges, although not entirely trivial. Indeed, IRS-III SWS integration allows the description of *any* XML data source available on the Web. From an expert point of view, the data source integration approach presents notable advantages compared to approaches based on standards such as the one demonstrated in the OWS-3 Initiative.[5] These advantages are framework openness (i.e., standards make integration easier but are not mandatory) and high-level service support (i.e., all the benefits of the underlying SWS platform, such as discovery, composition, etc. are immediately available). The steps involved in the process of adding a new data source, as well as the ability to automate each step, are described in the following:

- *ontological description of service:* the service, composed of the data types involved as well as its interface, can be described in a low-level ontology, i.e., at a level to remain close to the data. This step can be automated in many cases based on information contained in the schema of the service.
- *lifting definition*: the lifting operation allows the passage of data-type instances from a syntactic level (XML) defined in the data schema to an ontological one (OCML) specified in the ontology definition. This process can be automated every time the previous step is.
- *goal description*: a new goal has to be defined that represents the newly integrated Web service.
- *mediator description*: the goal has to be linked to the WS with a mediator, which is often a trivial operation.

- *lowering definition*: the lowering operation transforms instances of aggregation ontologies into syntactic documents to be used by the server and client applications. It is automatic since integration ontologies do not change.
- *mapping to integration ontologies*: this process is achieved by the knowledge engineer who modifies an ontology, defining which affordances are relevant to which context, with immediate effect.

This last step, which links affordances to a context rather than to a map, a system or simply to an object, allows meaning to emerge into an otherwise overwhelming amount of geographic data. This is absent from automatic syntactic mash-up builders,[117] or even from semantic ones such as *Geo-Names,*[118] which gather feeds on a map without taking context into account.

Other similar approaches that seem context-aware, such as the use of *tasks* in the recent – and mostly undisclosed at the time of writing – ESRI *ArcGIS Explorer* product,[119] are, to the best of our knowledge, actions attached to maps and that return heterogeneous results. These results do not seem to provide new tasks in a uniform and meaningful way.

An alternative method of adding meaning to spatial data can be found in the AKTive.Response[120] approach, where the Compendium tool is used for *collective sensemaking*, i.e., while gathering information from multiple (spatial) data sources collectively building context-relevant concept maps on the fly, with the help of other various ontology-aware tools (Tate et al. 2006). However, context, task relevance and the choice of affordances are still mostly left to the emergency planner, and data source integration seems to include essentially information messages.

25.6 Conclusions

The eMerges approach to spatial data integration presents advantages for the end user as well as for the data integration expert. Indeed, it allows the end user to handle tasks in a data-rich environment without being overwhelmed by the amount of information or by the complexity of the queries, and to the expert an easier approach to data integration.

In 2006, the eMerges prototype won a prize for the integration of Web scripting technologies with Semantic Web ones[121] and was selected among the five finalists of the *Semantic Web Challenge.*[122] Future developments will include an increase in the complexity of the integration ontologies (spatial, HCI and archetypes) in order to allow multirepresentation and an improved management of context to offer more cognitively sound features. Also, making the integration of new data sources even easier constitutes a long-term goal for the IRS SWS execution platform.

Bibliography

Abdelmoty, A.I., Smart, P.D., Jones, C.B., Fu, G. and Finch, D. (2005). "A Critical Evaluation of Ontology Languages for Geographic Information Retrieval on the Internet", *Journal of Visual Languages and Computing,* 16(4): 331–358.

Abolhassani, M., Fuhr, N. and Govert, N. (2003). "Information Extraction and Automatic Markup for XML Documents", *Intelligent Search on XML Data*. Eds. H. Blanken et al. Berlin: Springer, 159–174.

Agarwal, P., Bera, R. and Claramunt, C. (2006). "A Social and Spatial Network Approach to the Investigation of Research Communities over the World Wide Web", *4th International Conference on Geographic Information Science, Münster, Germany (LNCS, Vol 4197).* Eds. H. Miller et al. Berlin: Springer, 1–17.

Ahuja, S., Carriero, N. and Gelernter, D. (1986). "Linda and Friends", *IEEE Computer*, 19(8): 26–34.

AirBase (2005). *The European Air Quality Measurements Database*, EEA-ETC/ACC. http://air-climate.eionet.europa.eu/databases/airbase/.

Aitken, S.C. (2002). "Public Participation, Technological Discourses and the Scale of GIS", *Community Participation and Geographic Information Systems.* Eds. W. J. Craig et al. New York: Taylor and Francis, 357–366.

Aleman-Meza, B., Halaschek-Wiener, C., Budak Arpinar I., Ramakrishnan, C. and Sheth, A. (2005). "Ranking Complex Relationships on the Semantic Web", *IEEE Internet Computing*, 9(3): 37–44.

Amann, M., Bertok, I., Cofala, J., Gyarfas, F., Heyes, C., Klimont, Z., Schöpp, W. and Winiwarter, W. (2005)."Baseline Scenarios for the Clean Air for Europe (CAFE) Programme", *CAFE Scenario Analysis Report No. 1.* Laxenburg: International Institute for Applied Systems Analysis. http://ec.europa.eu/environment/-air/cafe/activities/pdf/cafe_scenario_report_1.pdf.

Amitay, E., Har'El, N., Sivan, R. and Soffer, A. (2004). "Web-a-Where: Geotagging Web Content", *27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* Sheffield, UK: ACM Press, 273–280.

Andrews, K., Guetl, C., Moser, J., Sabol, V. and Lackner, W. (2001). "Search Result Visualisation with xFIND", *Second International Workshop on User Interfaces to Data Intensive Systems (UIDIS 2001).* Zurich: IEEE Press, 50–58.

Annoni, A., Atkinson, M., Denzer, R., Hecht, L., Millot, M., Pichler, G., Sassen, A.-M., Couturier, M., Alegre, C., Sassier, H., Coene, Y. and Marchetti, P. G. (2005a). "Towards an Open Disaster Risk Management Service Architecture for INSPIRE and GMES", *European Commission*, (9). http://www.eu-orchestra.org/docs/20050223_White Paper_v9.pdf.

Annoni, A., Bernard, L., Douglas, J., Greenwood, J., Laiz, I., Lloyd, M., Sabeur, Z., Sassen, A.-M., Serrano, J. and Usländer, T. (2005b). "Orchestra: Developing a Unified Open Architecture for Risk Management Applications", *1st International Symposium on Geo-Information for Disaster Management, Delft, The Netherlands*. Ed. P. van Oosterom. Berlin: Springer, 1–18.

Anyanwu, K. and Sheth, A. (2003). "The ρ Operator: Discovering and Ranking Associations on the Semantic Web", *12th International World Wide Web Conference (WWW-2003),* Budapest, Hungary, 690–699.

Arpinar, B., Sheth, A., Ramakrishnan, C., Usery, L., Azami, M. and Kwan, M.-P. (2006). "Geospatial Ontology Development and Semantic Analytics", *Transactions in GIS,* 10(4): 551–575.

Arroyo, S., Lara, R., Gomez, J.M., Berka, D., Ding, Y. and Fensel, D. (2004). "Semantic Aspects of Web Services", *Practical Handbook of Internet Computing.* Ed. M. P. Singh. Baton Rouge, LA: Chapman Hall and CRC Press, 31/1–31-17.

Assfalg, R., Goebels, U. and Welter, H. (1998). *Internet Datenbanken, Konzepte, Methoden, Werkzeuge*. Bonn: Addison-Wesley-Longmann.

Auto-Oil II Programme: Air Quality Report (2000). "Final Version 7.1. Report of the Working Group 1 on Environmental Objectives", *Directorate General of the Environment*. http://autooil.jrc.cec.eu.int/finalaq.htm.

Baccino, T., and Pynte, J. (1994). "Spatial Coding and Discourse Models during Text Reading", *Language and Cognitive Processes*, 9(2): 143–155.

Baeza-Yates, R.A. and Ribeiro-Neto, B.A. (1999). *Modern Information Retrieval.* Harlow: ACM Press/Addison-Wesley.

Baldzer, J., Boll, S., Klante, P., Krösche, J., Meyer, J., Rump, N. and Scherp, A. (2004). "Location-Aware Mobile Multimedia Applications on the Niccimon Platform", *2. Braunschweiger Symposium - Informationssysteme für mobile Anwendungen (IMA-2004),* Brunswick, Germany, 318–334.

Batty, M. (2004). "Distance in Space Syntax", *Working Paper, no. 80*. Centre for Advanced Spatial Analysis (UCL). London.

Bebout, B.M., Carpenter, S.P., Des Marais, D.J., Discipulo, M., Embaye, T., Garcia-Pichel, F., Hoehler, T.M., Hogan, M., Jahnke, L.L., Keller, R.M., Miller, S.R., Prufert-Bebout, L.E., Raleigh, C., Rothrock, M. and Turk, K. (2002). "Long Term Manipulations of Intact Microbial Mat Communities in a Greenhouse Collaboratory: Simulating Earth's Present and Past Field Environments", *Astrobiology*, 2: 383–402.

Beier, R. and Doppelfeld, A. (1999). "Spatial Interpolation and Representativeness of Air Quality Data – an Intuitive Approach", *International Conference – Air Quality in Europe: Challenges for the 2000s*. Venice, Italy.

Bekkerman, R. and McCallum, A. (2005). "Disambiguating Web Appearances of People in a Social Network", *14th International Conference on World Wide Web*. Chiba, Japan: ACM Press, 463–470.

Bell, B. (2006). Vienna Marking Mozart Milestone. *BBC News*. London. http://news.bbc.co.uk/2/hi/entertainment/4654880.stm.

Benjamins, R., Contreras, J., Corcho, O. and Gómez-Pérez, A. (2004). "Six Challenges for the Semantic Web", *AIS SIGSEMIS Bulletin,* 1(1): 24–25.

Bennet, B. (2001). "Application of Supervaluation Semantics to Vaguely Defined Spatial Concepts, Spatial Information Theory: Foundations of Geographic Information Science", *Conference on Spatial Information Theory, Morro Bay, CA (LNCS, Vol. 2205).* Ed. D.R. Montello. New York: Springer, 108–123.

Bennett, B. (1996). "The Application of Qualitative Spatial Reasoning to {GIS}", *1st International Conference on GeoComputation.* Ed. R.J. Abrahart. Leeds, UK, 44–47.

Béra, R. and Claramunt, C. (2003). "Relative Adjacencies in Spatial Pseudo-partitions", *Conference on Spatial Information Theory (COSIT-2003), Ittingen, Switzerland (LNCS, Vol. 2825).* Eds. W. Kuhn et al. 218–234.

Bernard, L., Einspanier, U., Haubrock, S., Hubner, S., Kuhn, W., Lessing, R., Lutz, M. and Visser, U. (2003). "Ontologies for Intelligent Search and Semantic Translation in Spatial Data Infrastructures", *Photogrammetrie – Fernerkundung – Geoinformation,* 6: 451–462.

Berners-Lee, J., Hendler, J. and Lassila, O. (2001). "The Semantic Web", *Scientific American,* 184(5): 34–43.

Berrios, D.C., Sierhuis, M. and Keller, R.M. (2004). "Organizational Memory for Interplanetary Collaborative Scientific Investigations", *7th International Mars Society Conference.* Chicago, IL.

Biever, C. (2005). "Will Google Help Save the Planet?" *New Scientist,* 187(2512): 28–29.

Bilhaut, F. (2003). "The Linguastream Platform", *19th Spanish Society for Natural Language Processing Conference (SEPLN-2003)*. Madrid, Spain. http://www.sepln.org/revistaSEPLN/revista/31/31-Pag339.pdf., 339–340.

Bill, R. (1999). "Grundlagen der Geoinformationssysteme", *Bd. 1 – Hardware, Software und Daten*. Heidelberg: Wichmann Verlag.

Blond, N., Bel, L. and Vautard, R. (2003). "Three-Dimensional Ozone Data Analysis with an Air Quality Model over the Paris Area", *Journal of Geophysical Research*, 108(D23): 4744.

Marcheggiani E., Bocci, M., Colantonio, R. Galli, A. (2007). "Synthetic Indicators for an Integrated Landscape Planning Model", Methods and Instruments for the Scientifically Sound Use of Landscape and Socio-Ecological Indicators in Planning, *Landscape Online.* http://www.landscapeonline.de. In Press.

Bordin, P. (2002). *SIG: Concepts, Outils et Données*. London: Hermes Science.

Borillo, A. (1998). L'espace et son Expression en Français. Paris: Ophrys.

Botts, M. (2005). "OpenGIS Sensor Model Language (SensorML)", *Document 05-086*. Wayland, MA. http://portal.opengeospatial.org/files/?artifact id=12606.

Botts, M., Robin, A., Davidson, J. and Simonis, I. (2006). "OpenGIS Sensor Web Enablement Architecture Document", *Open Geospatial Consortium, Inc.* http://www.opengeospatial.org/about/?page=ipr.

Brandes, U. (2001). "A Faster Algorithm for Betweenness Centrality", *Journal of Mathematical Sociology*, 25(2): 163–177.

Bravo, J., Hervas, R., Chavira, G. and Nava, S. (2006). "Modeling Contexts by RFID-Sensor Fusion", *4th International IEEE Conference on Pervasive Computing and Communications Workshops (PerCom-2006)*. Pisa, Italy, 30–34.

Briggs, D.J., de Hoogh, C., Gulliver, J., Wills, J., Elliott, P., Kingham, S. and Smallbone, K. (2000). "A Regression-based Method for Mapping Traffic-related Air Pollution: Application and Testing in Four Contrasting Urban Environments", *The Science of the Total Environment,* 253: 151–167.

Brill, E. (1992). "A Simple Rule-based Part of Speech Tagger", *3rd Conference on Applied Computational Linguistics.* Trentino, Italy, 152–155.

Brodlie, K.W., Carpenter, L.A., Earnshaw, R.A., Gallop, J.R., Hubbold, R.J., Mumford, A.M., Osland, C.D. and Quarendon, P. (1991). *Scientific Visualization: Techniques and Applications*. London: Springer.

Bruns, T.H. and Egenhofer, M.J. (1996). "Similarity of Spatial Scenes", *7th International Symposium on Spatial Data Handling (SDH-1996)*. Eds. M.-J. Kraak and M. Molenaar. Delft, Netherlands, 31–42.

Burrough, P. and Frank, A. (1995). "Concepts and Paradigms in Spatial Information: Are Current Geographical Information Systems Truly Generic?" *International Journal of Geographical Information Systems*, 9(2): 101–116.

Busch, D. (2005). "Datenbank-Managementsysteme", *Umweltinformationssysteme.* Ed. P. Fischer-Stabel. Heidelberg: Wichmann Verlag, 119–130.

Buse, A., Norris, D., Harmens, H., Büker, P., Ashenden, P. and Mills, G. (2003). "Heavy Metals in European Mosses", *2000/2001 Survey*, Bangor, UK: UNECE ICP Vegetation.

Butler, D. (2006). "Virtual Globe: The Web-Wide World", *Nature*, 439: 776–778.

Bytnerowicz, A., Godzik, B., Fraczek, W., Grodzinska, K., Krywult, M., Badea, O., Barancok, P., Blum, O., Cerný, M. and Godzik, S. (2002). "Distribution of Ozone and Other Air Pollutants in Forests of the Carpathian Mountains in Central Europe", *Environmental Pollution,* 116: 3–25.

Cabral, L., Domingue, J., Galizia, S., Gugliotta, A., Norton, B., Tanasescu, V. and Pedrinaci, C. (2006). "IRS-III: A Broker for Semantic Web Services based Applications", *5th International Semantic Web Conference (ISWC-2006)*. Eds. I. Cruz et al. Athens, GA: Springer, 201–214.

Cai, M. and Frank, M. (2004) "RDFPeers: A Scalable Distributed RDF Repository based on a Structured Peer-to-Peer Network", *World Wide Web Conference (WWW-2004)*. New York.

Campbell, N., Muller, H. and Randell, C. (1999) "Combining Positional Information with Visual Media", *3rd International Symposium on Wearable Computers.* Ed. The IEEE Computer Society, 203–205.

Cannon, H.M. (2001). "Addressing New Media with Conventional Media Planning", *Journal of Interactive Advertising,* 1(2). http://www.jiad.org/vol1/no2/.

Cargill, C. (1997). "Sun and Standardization Wars," *ACM StandardView,* 5(4): 133–135.

Castelli, G., Rosi, A., Mamei, M. and Zambonelli, F. (2006). "Browsing the World: Bridging Pervasive Computing and the Web", *International Workshop on Ubiquitous Information Systems.* Münster, Germany.

Cayzer, S. and Butler, M. (2004). "Semantic Photos", *Hewlett Packard Labs Technical Report.* http://www.hpl.hp.com/techreports/2004/HPL-2004-234.html.

Cazenave, J., Marquesuzaà, C., Dagorret, P. and Gaio, M. (2004). "La Revitalisation Numérique du Patrimoine Littéraire Territorialisé", *Colloque International EBSI-ENSSIB*. Montréal. http// :www.ebsi.umontreal.ca/rech/ebsi-enssib/pdf/casenave_et_al.pdf.

CEA (2005). "Flood Prevention in Europe. The Role of the Insurance Industry", *Property Insurance Committee*. http://www.cea.assur.org/cea/v1.1/actu/-pdf/uk/annexe232.pdf.

CGDI (2005). *CGDI Architecture Description, Version 2.0.* Canadian Geospatial Data Infrastructure Architecture Working Group. http://www.geoconnections.org/-publications/tvip/arch_E/CGDI_Architecture_final_E.html.

Chakrabarti, S., van den Berg, M. and Domc, B. (1999). "Focused Crawling: A New Approach to Topic-specific Web Resource Discovery." *Computer Networks,* 31(11–16): 1623–1640.

Chalmers, M. (1993). "Using a Landscape Metaphor to Represent a Corpus of Documents", *Spatial Information Theory: A Theoretical Basis for GIS (LNCS, Vol. 716)*. Eds. A.U. Frank and I. Campari. Berlin: Springer, 377–390.

Chalmers, M. (1996). "A Linear Iteration Time Layout Algorithm for Visualising High-Dimensional Data", *7th Conference on Visualization*. San Francisco, CA: IEEE Computer Society, 127–132.

Chang, Y.S. and Park, H.-D. (2006). "XML Web Service-based Development Model for Internet GIS Applications", *International Journal of Geo-Information Science*, 20(4): 371–399.

Chappell, A. and Jewell, T. (2002). *Java Web Services.* Sebastopol, CA: O'Reilly.

Charnois, T., Mathet, Y., Enjalbert, P. and Bilhaut, F. (2003). "Geographic Reference Analysis for Geographic Document Querying", *Workshop on the Analysis of Geographic References, Human Language Technology Conference (NAACL/HLT-2003)*. Association for Computational Linguistic, Edmonton, Canada. http://people.mokk.bme.hu/~kornai/NAACL/WS9/ws904.pdf.

Chaze, X. and Napoli, A. (2004). "Conception et Développement d'un Portail Internet sur les Risques Naturels en France à l'Aide de la Technologie Map-Server", *Final Internship Report for National School of Geographical Science (ENSG) in Partnership between Pôle Cindyniques,* 80.

Chin, G.J. and Lansing, C.S. (2004). "Capturing and Supporting Contexts for Scientific Data Sharing via the Biological Sciences Collaboratory", *ACM Conference on Computer Supported Cooperative Work (CSCW-2004)*. Chicago: ACM Press, 409–418.

Chinchor, N. (1997). "MUC-7 Named Entity Task Definition", *7th Message Understanding Conference*. Fairfax, VA. http://www-nlpir.nist.gov/related_projects/-muc/.

Chirita, P.A., Idreos, S., Koubarakis, M. and Nejdl, W. (2004). "Publish/Subscribe for RDF-based P2P Networks", *1st European Semantic Web Symposium (ESWS-2004)*. Eds. C. Bussler et al. Heraklion: Springer, 182–197.

Choicki, J. (1999). "Constraint-based Interoperability of Spatiotemporal Databases". *Geoinformatica,* 3(3): 211–243.

Chong, C.Y. and Kumar, S.P. (2003). "Sensor Networks: Evolution, Opportunities, and Challenges", *Proceedings of the IEEE*, 91(8): 1247–1256.

Chrisman, N.R. (1987). "Design of Geographic Information Systems based on Social and Cultural Goals", *Photogrammetric Engineering and Remote Sensing,* 53(10): 1367–1370.

Clancey, W.J., Sachs, P., Sierhuis, M. and Hoof, R.V. (1998). "Brahms: Simulating Practice for Work Systems Design", *International Journal of Human-Computer Studies,* 49: 831–865.

Clancey, W.J., Sierhuis, M., Alena, R., Berrios, D., Dowding, J., Graham, J. S., Tyree, K.S., Hirsh, R.L., Garry, W.B., Semple, A., Buckingham Shum, S.J., Shadbolt, N. and Rupert, S.M. (2005). "Automating CapCom Using Mobile Agents and Robotic Assistants", *1st Space Exploration Conference.* Orlando, FL. http://eprints.aktors.org/375/.

Clancey, W.J., Sierhuis, M., Kaskiris, C. and Hoof, R.V. (2002). "Brahms Mobile Agents: Architecture and Field Tests", *Fall Symposium on Human-Robot Interaction (AAAI-2002).* North Falmouth, MA: AAAI Press, 25–29.

Clementini, E., Sharma, J. and Egenhofer, M. (1994). "Modeling Topological Spatial Relations: Strategies for Query Processing", *Computers and Graphics,* 815–822.

Clinton, W.J. (1994). *Executive Order 12906 – Coordinating Geographic Data Acquisition and Access: The National Spatial Data Infrastructure.* Washington: The White House. http://www.fas.org/irp/offdocs/eo12906.htm

Clinton, W.J. (2006). "UN Special Envoy for Tsunami Recovery", *Third International Conference on Early Warning (EWC-2006).* Bonn, Germany.

Clough, P. (2005). "Extracting Metadata for Spatially-Aware Information Retrieval on the Internet", *2nd International Workshop on Geographic Information Retrieval (GIR-2005).* Bremen, Germany, 25–30.

Cohn, A.G. and Gotts, N.M. (1996). "The 'Egg-Yolk' Representation of Regions with Indeterminate Boundaries", *GISDATA Specialist Meeting on Spatial Objects with Undetermined Boundaries.* Eds. P. Burrough and A.M. Frank. London: Taylor and Francis, 171–187.

Cohn, A.G., Bennett, B., Gooday, J. and Gotts, N.M. (1997). "Qualitative Spatial Representation and Reasoning with the Region Connection Calculus", *GeoInformatica,* 1(3): 275–316.

Cole, S. and Hornsby, K. (2005). "Modeling Noteworthy Events in a Geospatial Domain", *1st International Conference on Geospatial Semantics (GeoS-2005), Mexico City, (LNCS, Vol 3799).* Springer, 77–89.

Couclelis, H. (1992). "People Manipulate Objects (but Cultivate Fields): Beyond the Raster-Vector Debate in GIS", *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space.* Eds. I. Frank et al. New York: Springer, 65–77.

Cowie, J. and Lehnert, W. (1996). "Information Extraction", *Communications of the ACM,* 39(1): 80–91.

Cox, S. (2005). "Observations and Measurements". *Document 05-087r5.* Wayland, MA. http://portal.opengeospatial.org/files/?artifact id=14034.

Cox, S., Daisey, P., Lake, R., Portele, C. and Whiteside, A. (2003). "Geography Markup Language (GML) 3.0", *OpenGIS Specification*. http://www.opengis.net/gml/.

Crampton, J. (1995). "The Ethics of GIS", *Cartography and Geographic Information Systems*, 22(1): 84–89.

Cressie, N. (1993). *Statistics for Spatial Data*. New York: John Wiley and Sons.

CTG (2001). "Sharing the Costs, Sharing the Benefits: The NYS GIS Cooperative Project", *Project Report 95-4.* Center for Technology in Government. http://www.ctg.albany.edu/publications/reports/sharing_the_costs.

Cunningham, H., Wilks, Y. and Gaizauskas, R. (1996). "GATE – a General Architecture for Text Engineering". *16th Conference on Computational Linguistics (COLING-1996)*. Copenhagen. http://citeseer.comp.nus.edu.sg/cunningham96gate.html.

Curino, C., Giani, M., Giorgetta, M., Giusti, A., Murphy, A.L. and Picco, G.P. (2005). "Mobile Data Collection in Sensor Networks: The TinyLIME Middleware", *International Conference on Pervasive Computing and Communications (PerCom-2005)*. Kona, HI, 446–469.

Cutting D. (2006). "Apache Lucene, a High-Performance, Full-featured Text Search Engine Library Written Entirely in Java", *Apache Software Foundation*. http://lucene.apache.org/.

d'Amato, C., Fanizzi, N. and Esposito, F. (2005). "A Semantic Dissimilarity Measure for Concept Descriptions in Ontological Knowledge Bases", *2nd International Workshop on Knowledge Discovery and Ontologies (KDO-2005)*. Porto, Portugal. https://webhosting.vse.cz/svatek/KDO05/paper2.pdf.

Daviel, A. and Kaegi, F.A. (2003). Geographic Registration of HTML Documents (IETF Internet Draft, July 2003). Sterling: Internet Engineering Task Force. http://ecotroph.net/geopriv.

Davis, M. (2002). "JTS Topology Suite". http://www.vividsolutions.com/jts/jtshome.htm.

Dawson, F. and Howes, T. (1998). "vCard MIME Directory Profile (RFC2426)". http://www.ietf.org/rfc/rfc2426.txt.

de Berg, M., van Kreveld, M., Overmars, M. and Schwarzkopf, O. (2000). "Computational Geometry: Algorithms and Applications", 2nd Edition. Berlin: Springer.

de Bruijn, J. (2005). "The Web Service Modeling Language". *Final Draft D16.v0.21*. http://www.wsmo.org/TR/d16/d16.1/v0.21/.

de la Asuncion, M., Castillo, L., Fdez.-Olivares, J., Garca-Prez, O., Gonzlez, A. and Palao, F. (2005). "SIADEX: An Interactive Artificial Intelligence Planner for Decision Support in Forest Fire Fighting", *Artificial Intelligence Communications* (AIComm). Amsterdam: IOS Press, 18(4): 257–268.

Dean, M. and Kolas, D. (2005). *Semwebcentral: Project Info – Geospatial Semantic Web.* http://project.semwebcentral.org/projects/gsw/.

Delboni, T.M., Borges, K.A.V. and Laender, A.H.F. (2005). "Geographic Web Search Based on Positioning Expressions", *2nd International Workshop on Geographic Information Retrieval (GIR-2005)*. Bremen, Germany, 61–64.

Denby, B., Horálek, J., Walker, S.E., Eben, K. and Fiala, J. (2005). "Interpolation and Assimilation Methods for European Scale Air Quality Assessment and Mapping Part I: Review and Recommendations", *ETC/ACC Technical Paper 2005/7*. http://air-climate.eionet.europa.eu/reports/.

DeVarco, B. (2004). "Earth as a Lens: Global Collaboration, GeoCommunication, and the Birth of EcoSentience", *PlaNetwork Journal*. http://journal.planetwork.net/article.php?lab=devarco0704.

Devillers, R., Gervais, M., Bedard, Y. and Jeansoulin, R. (2003). "Spatial Data Quality: From Metadata to Quality Indicators and Contextual End-User Manual", *OEEPE/ISPRS Joint Workshop on Spatial Data Quality Management*. Istanbul, Turkey.

Dey, A., Abowd, G.D. and Salber, D. (2001). "A Conceptual Framework and a Toolkit for Supporting the Rapid Prototyping of Context-Aware Applications", *Human-Computer Interaction*, 16: 97–166.

Dey, A.K. and Abowd, G.D. (2000). "Towards a Better Understanding of Context and Context-Awareness". *Workshop on the What, Who, Where, When and How of Context-Awareness, Conference on Human Factors in Computer Systems (CHI-2000)*. New York: ACM Press.

Dey, A.K., Abowd, G.D. and Salber, D. (1999). "A Context-based Infrastructure for Smart Environments", *International Workshop on Managing Interactions in Smart Environments*. Dublin, Ireland.

Dickmann, F. (2001). "Compass – Das Geographische Seminar, Webmapping und Web-Gis", Braunschweig: Westermann Verlag.

Diligenti, M., et al. (2000). "Focused Crawling Using Context Graphs", *26th International Conference on Very Large Databases (VLDB-2000)*. Cairo, Egypt.

Ding, J., Gravano, L. and Shivakumar, N. (2000). "Computing Geographical Scopes of Web Resources", *26th International Conference on Very Large Data Bases*. Cairo: Morgan Kaufmann, 545–556.

Dolbear, C. and Hart, G., (2006). "R2D2: Combining Spatial and Semantic Queries into Spatial Databases", *Technical Paper*. http://www.ordnancesurvey.co.uk/research.

Domingue, J. and Motta, E. (2000). "PlanetOnto: From News Publishing to Integrated Knowledge Management Support", *IEEE Intelligent Systems,* 15(3): 26–32.

Dong, X., Halevy, A. and Madhavan, J. (2005). "Reference Reconciliation in Complex Information Spaces", *24th International Conference on Management of Data (SIGMOD-2005)*. Baltimore, MD: ACM Press, 85–96.

Doyle, A. and Reed, C. (2001). "Introduction to OGC Web Services", *Open Geospatial Consortium White Papers*. http://www.opengeospatial.org/pressroom/papers.

Dunne, J.A., Williams, R.J. and Martinez, N.D. (2002). "Food Web Structure and Network Theory, the Role of Connectance and Size", *National Academy of Sciences*, USA, 99(20): 12917–12922.

Dymon, U.J. (2003). "An Analysis of Emergency Map Symbology", *International Journal of Emergency Management,* 1(3): 23–27.

EC (1996). Air Quality Framework Directive, Council Directive 96/62/EC on Ambient Air Quality Assessment and Management, *Official Journal of the European Communities (OJ),* L 296, 21.11.1996, 55–63. http://eur-lex.europa.eu/LexUriServ/site/en/consleg/1996/L/01996L0062-20031120-en.pdf.

EC (1997). Council Decision 97/101/EC, Establishing a Reciprocal Exchange of Information and Data from Networks and Individual Stations Measuring Ambient Air Pollution within the Member States, (*OJ) L 035, 05.02.1997, 14-22,* and Its Amended Annexes to 97/101/EC, *Commission Decision 2001/752/EC, (OJ) L 282, 26.10.2001, 69-76*. http://eur-lex.europa.eu/.

EC (1999a). First Daughter Directive, Council Directive 1999/30/EC Relating to Limit Values for Sulphur Dioxide, Nitrogen Dioxide and Oxides of Nitrogen, Particulate Matter, and Lead in Ambient, *OJ L 163, 29.06.1999, 41-60.* http://eur-lex.europa.eu/.

EC (1999b). *EU Focus on Clean Air*, European Commission: http://ec.europa.eu/environment/eufocus/clean_air.pdf.

EC (2002). Third Daughter Directive, Council Directive 2002/3/EC Relating to Ozone in Ambient Air, *OJ L 67, 09.03.2002, 14-30*. http://eur-lex.europa.eu/.

EEA (2005). *Corine Land Cover 2000 (CLC2000) 250 m – Version 8/2005*. European Environmental Agency. Copenhagen, Denmark. http://dataservice.eea.europa.eu/.

Egenhofer, M.J. (2002). "Toward the Semantic Geospatial Web", 10th ACM International Symposium on Advances in Geographic Information Systems (*ACM-GIS-2002*). Virginia: ACM Press, 1–4.

Egenhofer, M.J. and Franzosa, R.D. (1991). "Point-Set Topological Relations", *International Journal for Geographic Information Sytems*, 5(2): 161–174.

Egenhofer, M.J. and Herring, J. (1995). *Advances in Spatial Databases. Fourth International Symposium (SSD-1995)*, Portland, ME.

Egenhofer, M.J. and Mark, D.M. (1995). "Naive Geography", *Conference on Spatial Information Theory (COSIT-1995)*. Eds. A.U. Frank and W.A. Kuhn. Semmering, Austria: Springer, 1–15.

Eisenstadt, M., Komzak, J. and Dzbor, M. (2003). "Instant Messaging + Maps = Powerful Collaboration Tools for Distance Learning", *International Symposium on Tele Education* (*TelEduc-2003)*. Havana, Cuba. http://kmi.open.ac.uk/-projects/buddyspace/docs/eisenstadt-komzak-dzbor-teleduc03.doc.

Environmental Systems Research Institute, Inc. (1991). "Cell-based Modeling with GRID", Redlands, CA: ESRI Press.

Erle, S. and Gibson, R. (2006). *Google Maps Hacks,* Sebastopol, CA: O'Reilly Press.

Erle, S., Gibson, R. and Walsh, J. (2005). *Mapping Hacks – Tips & Tools for Electronic Cartography*. Sebastopol, CA: O'Reilly Press.

Espinoza, F., Persson, P., Sandin, A., Nystrom, H., Cacciatore, E. and Bylund, M. (2001). "GeoNotes: Social and Navigational Aspects of Location-based Information Systems", *International Conference on Ubiquitous Computing (UbiComp-2001)*. Atlanta, GA: Springer, 2–17.

ESRI (2004). "ArcIMS 9.0 Architecture and Functionality", *ESRI White Paper*. ESRI Press. http://www.esri.com/library/whitepapers/pdfs/ arcims9-architecture.pdf.

ESRI (2005). "Altitude, 30x30 grid, GTOPO30", *Global Digital Elevation Model*. Redlands, CA: ESRI Press.

Etcheverry, P., Marquesuzaà, C. and Corbineau S. (2006). "Designing Suited Interactions for a Document Management System Handling Localized Documents", *24th ACM International Conference on Design of Communication (SIGDOC-2006)*. Myrtle Beach, SC, 188–195.

Etcheverry, P., Marquesuzaà, C. and Lesbegueries, J. (2005). "Revitalisation de Documents Territorialisés : Principes, Outils et Premiers Résultats", *Workshop Met-SI INFORSID*. Grenoble, France.

Ewalt, D.M. (2005). "Google Is Everywhere", *Forbes*. http://www.forbes.com/technology/2005/09/02/hurricane-google-map-rescue-cx_de_0902google.html.

Fagerli, H., Simpson, D. and Tsyro, S. (2004). "Unified EMEP Model: Updates", *EMEP Status Report 1/2004*. Oslo. http://www.emep.int/publ/reports/2004/Status_report_int_del1.pdf.

Falke, S.R. and Husar, R.B. (1998a). "Declustering in the Spatial Interpolation of Air Quality Data". *90th Air and Waste Management Association Annual Meeting, CAPITA Paper* No. 98-A927. St. Louis, MO. http://capita.wustl.edu/CAPITA/Awma98/HTTP/98_A927.htm.

Falke, S.R. and Husar, R.B. (1998b). "Maps of PM2.5 over the U.S. Derived from Regional PM2.5 and Surrogate Visibility and PM10 Monitoring Data", *CAPITA Paper* No. 98-A918. St. Louis, MO. http://capita.wustl.edu/CAPITA/Awma98/HTTP/98_A918.htm.

Farrukh Najmi (2006). "Web Content Management Using the OASIS ebXML Registry Standard". http://ebxmlrr.sourceforge.net/presentations/xmlEurope2004/04-02-02.pdf.

Federal Geographic Data Committee (1999). *National Spatial Data Clearinghouse Web Site*. http://clearinghouse1.fgdc.gov/.

Feier, C. and Domingue, J. (2005). *WSMO Primer*. http://www.wsmo.org/TR/d3/d3.1/v0.1/20050401/.

Fensel, D. and Bussler, C. (2002). "The Web Service Modeling Framework WSMF", *Electronic Commerce Research and Applications*, 1(2): 113–137.

Fidler, R. (1997). *Mediamorphosis: Understanding New Media*. Thousand Oaks, CA: Pine Forge Press.

FIPA (2002). "FIPA Abstract Architecture Specification", *Technical Specification SC00001L, Foundation for Intelligent Physical Agents*. Geneva, Switzerland. http://www.fipa.org.

Fonseca, F. and Sheth, A. (2002). "The Geospatial Semantic Web". *UCGIS White Paper*. http://www.ucgis4.org/priorities/research/2002researchagenda.htm.

Freeman, J. (1975). "The Modelling of Spatial Relations", *Computer Graphics and Image Processing*, 4: 156–171.

Freeman, L.C. (1977). "A Set of Measures of Centrality Based on Betweenness", *Sociometry*, 40: 35–41.

Friedmannová, L., Konecný, M. and Stanek, K. (2006). "An Adaptive Cartographic Visualization for Support of the Crisis Management", *Autocarto 2006*. Vancouver, Canada. http://www.cartogis.org/publications.

Frost, P. and Vosloo, H. (2006). "Providing Satellite-based Early Warnings of Fires to Reduce Fire Flashovers on South African Transmission Lines", *10th Biennial Australasian Bushfire Conference (Bushfire-2006),* Brisbane, Australia. http://www.fireandbiodiversity.org.au/.

Fuger, S., Farrukh, N. and Stojanovic, N. (2005). "Oasis Standard: ebXML Registry Information Model Version 3.0". http://docs.oasis-open.org/regrep-rim/v3.0/regrep-rim-3.0-os.

Gahegan, M. (1995). "Proximity Operators for Qualitative Spatial Reasoning", *Spatial Information Theory*: *International Conference (COSIT-1995)*. Semmering, Austria: Springer, 31–44.

Gahleitner, E., Behrendt, W., Palkoska, J. and Weippl, E. (2005). "Knowledge Sharing and Reuse: On Cooperatively Creating Dynamic Ontologies", *16th ACM Conference on Hypertext and Hypermedia (Hypertext-2005)*. Salzburg, Austria: ACM Press.

Gaizauskas, R. (2002). "An Information Extraction Perspective on Text Mining: Tasks, Technologies and Prototype Applications", *Euromap Text Mining Seminar*. Sheffield, UK.

Gaizauskas, R. and Wilks, Y. (1998). "Information Extraction: Beyond Document Retrieval ", *Journal of Documentation,* 70–105.

Gaizauskas, R., Wakao, T., Humphreys, K., Cunningham, H. and Wilks, Y. (1995). *University of Sheffield: Description of the Lasie System as Used for Muc.* http://acl.ldc.upenn.edu/M/M95/M95-1017.pdf.

Galdos Systems Inc. (2003). "Developing and Managing GML Application Schemas: Best Practices". http://www.geoconnections.org/developersCorner/devCorner_devNetwork/components/GML_bpv1.3_E.pdf.

Gale, W., Church, K. and Yarowsky, D. (1992). "One Sense per Discourse", *5th DARPA Speech and Natural Language Workshop*. New York: Harriman, 233–237.

Galton, A. and Hood, J. (2005). "Anchoring: A New Approach to Handling Indeterminate Location in GIS", *Spatial Information Theory: International Conference (COSIT-2005).* Eds. A.G. Cohn and M. David. New York: Springer, 1–13.

Galton, A.P. (2001). "A Formal Theory of Objects and Fields", *Conference on Spatial Information Theory (COSIT-2001)*. Ed. D.R. Montello. Berlin: Springer, 458–473.

Gao, W., Lee, H. C., and Miao, Y. (2006). "Geographically Focused Collaborative Crawling", *15th International Conference on World Wide Web*. Edinburgh, Scotland: ACM Press, 287–296.

Garbin, E. and Mani, I. (2005). "Disambiguating Toponyms in News", *Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Vancouver, Canada, 363–370.

Gärdenfors, P. (2000). "Conceptual Spaces: The Geometry of Thought". Cambridge, MA: Bradford Books, MIT Press.

Gartner, G. (2004). "Location-based Mobile Pedestrian Navigation Services – the Role of Multimedia Cartography." *International Joint Workshop on Ubiquitous, Pervasive and Internet Mapping (UPIMap-2004)*. Tokyo, Japan. http:// - www.ubimap.net/upimap2004/html/papers/UPIMap04-B-03-Gartner.pdf.

Gartner, G. (2006). "From Mobile towards Ubiquitous Cartography – Trends in Contemporary Cartography. Presentation." *1st International Conference on Cartography and GIS*. Ed. T. Bandrova. Borovets, Bulgaria, 21–22.

GeoConnections (2001). "GeoConnections Discovery Portal Website". http://geodiscover.cgdi.ca/.

Gervais, M. (2003). "Pertinence d'un Manuel d'Instruction au Sein d'une Stratégie de Gestion du Risque Juridique Découlant de la Fourniture de Données Géographiques Numériques", *Rapport de thèse, Marne la Vallé University*. Laval University, Québec.

Getty Trust (2006). "The Getty Thesaurus of Geographic Names". http://www.getty.edu/research/conducting_research/vocabularies/tgn/.

Gibson, J.J. (1986). *The Ecological Approach to Visual Perception.* Mahwah, NJ: Lawrence Erlbaum Associates.

Goldstone, R.L. and Son, J. (2005). "Similarity", *Cambridge Handbook of Thinking and Reasoning*. Eds. K. Holyoak and R. Morrison. Cambridge: Cambridge University Press, 13–36.

Goodchild, M.F. (2004). "GIScience: Geography, Form and Process", *Annals of the Association of American Geographers*, 94(4): 709–714.

Goodwin, J. (2005). "What Have Ontologies Ever Done for Us – Potential Applications at a National Mapping Agency". http://www.mindswap.org/2005/OWLWorkshop/sub18.pdf.

Gore, A. (1998). "The Digital Earth: Understanding Our Planet in the 21st Century", *Lecture at the California Science Center (31 Jan 1998)*. Los Angeles: http://www.digitalearth.gov/VP19980131.html.

Gräf, J., Henrich, A., Lüdecke, V. and Schlieder, C. (2006). "Geografisches Information Retrieval", *Datenbank-Spektrum,* 6(18): 48–56.

Greco, G., Greco, S. and Zumpano, E. (2002). "A Stochastic Approach for Modelling and Computing Web Communities", *3rd International Conference on Web Information Systems Engineering*. Singapore: IEEE Press, 43–52.

Groot, R. and McLaughlin, J. (2000). *Geospatial Data Infrastructure: Concepts, Cases and Good Practice*. New York: Oxford University Press.

Gruber, T.R. (1993). "A Translation Approach to Portable Ontology Specifications", *Knowledge Acquisition*, 5(2): 199–220.

GSDI (2001). *Developing Spatial Data Infrastructures: The SDI Cookbook, Version 1.1.* Global Spatial Data Infrastructure Association. http://www.gsdi.org/pubs/-cookbook/index.html.

Guarino, N. (1997). "Understanding, Building and Using Ontologies", *International Journal of Human-Computer Studies,* 46(2–3): 293–310.

Guarnieri, F. and Garbolino, E. (2003). "Systèmes d'Information et Risques Naturels: Pratiques, Innovations Méthodologiques et Technologiques", *Joint work*. Ecole des Mines de Paris Press.

Hammond, B., Sheth, A. and Kochut, K. (2002). "Semantic Enhancement Engine: A Modular Document Enhancement Platform for Semantic Applications over Heterogeneous Content", *Real World Semantic Web Applications*. Eds. V. Kashyap and L. Shklar. Amsterdam: IOS Press. 29–49.

Hansen, B. (2006). "The GeoURL ICBM Address Server": http://geourl.org/.

Harris, T.M. and Weiner, D. (1996). "GIS and Society: The Social Implications of How People, Space and Environment Are Represented in GIS", *Scientific Report for the Initiative 19 Special Meeting*. NCGIA Report, 96-7.

Harris, T.M. and Weiner, D. (1998). "Empowerment, Marginalization, and 'Community-integrated' GIS", *Cartography and Geographic Information Systems,* 25(2): 67–76.

Harris, T.M. and Weiner, D. (2002). "Implementing a Community-integrated GIS", *Community Participation and Geographic Information Systems.* Eds. W.J. Craig et al. New York: Taylor and Francis, 246–258.

Harris, T.M., Alagan, R. and Rouse L.J. (2002). "Geo-Visualization Approaches to PGIS Decision Making in Environmental Impact Assessment", *1st Annual Conference on PPGIS.* Park Ridge, IL: URISA, 180–185.

Harris, T.M., Weiner, D., Warner, T.A. and Levin, R. (1995). "Pursuing Social Goals through Participatory Geographic Information Systems: Redressing South Africa's Historical Political Ecology", *Ground Truth: The Social Implications of GIS*. Ed. J. Pickles. New York: Guildford Press, 196–222.

Hau, J., Lee, W. and Darlington, J. (2005). "A Semantic Similarity Measure for Semantic Web Services in Web Service Semantics", *International World Wide Web Conf*erence Workshop (*WWW-2005)*, Chiba, Japan.

Hazaël-Massieux, D. and Connolly, D. (2006). "Gleaning Resource Descriptions from Dialects of Languages (GRDDL)", *W3C Draft*. http://www.w3.org/2004/01/rdxh/spec.

Herman, L. (2001). *W3C Semantic Web Activity*. http://w3.org/2001/sw.

Hernandez, N. (2005). "Ontologies pour l'Aide à l'Exploration d'une Collection de Documents", *Ingénierie des Systèmes d'Information*. Paris: Hermès Sciences, 10: 11–31.

Herpin U., Lieth H. and Markert B. (1995). "Monitoring der Schwermetallbelastung in der Bundesrepublik Deutschland mit Hilfe von Moosanalysen", Berlin: UBA-Texte, 31/95.

Hill, L. (1999). "Indirect Geospatial Referencing through Place Names in the Digital Library: Alexandria Digital Library Experience with Developing and Implementing Gazetteers", *62nd Annual Meeting of the American Society for Information Science*. Medford, NJ: ASIS, 57–69.

Hill, L. (2000). "Core Elements of Digital Gazetteers: Place Names, Categories, and Footprints", *4th European Conference on Research and Advanced Technology for Digital Libraries (ECDL-2000)*. Lisbon, Portugal: Springer, 280–290.

Hill, L.L., Frew, J. and Zheng, Q. (1999). "Geographic Names – the Implementation of a Gazetteer in a Georeferenced Digital Library", *D-Lib Magazine,* 5(1). http://www.dlib.org/.

Hof, R.D. (2005). "Mix, Match, and Mutate: "Mash-ups" – Homespun Combinations of Mainstream Services – Are Altering the Net", *Business Week,* 3942 (July 25): 72.

Hogan, P. and Kim, R. (2004). "NASA Planetary Visualization Tool", *American Geophysical Union Fall Meeting.* San Francisco, CA.

Hong, J. (2002). "The Context Fabric: An Infrastructure for Context-Aware Computing", *Conference on Computer Human Interaction.* Minneapolis, MN.

Horálek, J., KurFürst, P., Denby, B., De Smet, P., De Leeuw, F., Brabec, M. and Fiala, J. (2005). "Interpolation and Assimilation Methods for European Scale Air Quality Assessment and Mapping, Part II: Development and Testing New Methodologies", *ETC/ACC Technical Paper 2005/8.* http://air-climate.eionet.europa.eu/reports/.

Horrocks, I., Patel-Schneider, P.F. and Harmelen, F.V. (2003). "From SHIQ and RDF to OWL: The Making of a Web Ontology Language", *Journal of Web Semantics,* 1(1): 7–26.

Horrocks, I., Patel-Schneider, P.F., Boley, H., Tabet, S., Grosof, B. and Dean, M. (2004). "SWRL: A Semantic Web Rule Language Combining OWL and RuleML". *W3C Member Submission*. http://www.w3.org/Submission/SWRL/.

Hou, J. and Zhang, Y. (2002). "A Matrix Approach for Hierarchical Web Page Clustering Based on Hyperlinks", *3rd International Workshops on Web Information Systems Engineering.* Singapore: IEEE Press, 207–216.

Humphreys, K., Gaizauskas, R., Azzam, S., Huyck, C., Mitchell, B., Cunningham, H. and Wilks, Y. (1998). "University of Sheffield: Description of the LaSIE-II System as Used for MUC-7", *7th Message Understanding Conference.* Fairfax, VA. http://www-nlpir.nist.gov/related_projects/muc/.

ISO (2000). *ISO/IEC 15444-1:2000 (JPEG 2000 Encoding Specification)*. http://www.jpeg.org/jpeg2000/index.html.

ISO (2003). "International Standard ISO 19119, Geographic Information Services", *ISO/TC 211*.

ISO (2004). ISO/IEC 15444-9:2004 (JPEG 2000 Encoding Specification – Part 9, Interactivity Tools, APIs and Protocols). http://www.jpeg.org/jpeg2000/index.html.

Jackson, J. (2006). "'Neogeography' Blends Blogs with Online Maps", *National Geographic News*. http://news.nationalgeographic.com/news/2006/04/.

Jain, A.K., Murty, M.N. and Flynn, P.J. (1999). "Data Clustering: A Review", *ACM Computing Surveys,* 31(3): 264–323.

Janowicz, K. (2005). "Extending Semantic Similarity Measurement by Thematic Roles", *1st International Conference on GeoSpatial Semantics* (GeoS-2005). Berlin: Springer, 137–152.

Janowicz, K. (2006). "Sim-DL: Towards a Semantic Similarity Measurement Theory for the Description Logic ALCNR in Geographic Information Retrieval", *2nd International Workshop on Semantic-based Geographical Information Systems (SeBGIS-2006), OTM Workshops, Montpellier, France (LNCS, Vol. 4278)*. Eds. R. Meersman et al. Berlin: Springer, 1681–1692.

Jarrar, M. and Meersman, R. (2002). "Formal Ontology Engineering in the DOGMA Approach", *International Conference on Ontologies, Databases and Applications of Semantics (LNCS, Vol. 2519)*. Eds. R. Meersman and Z. Tari. Berlin: Springer, 1238–1254.

Jerrett, M., Arain, M.A., Kanaroglou, P., Beckerman, B., Crouse, D., Gilbert, N.L., Brook, J.R., Finkelstein, N. and Finkelstein, M.M. (2003). "Modelling the Intra-Urban Variability of Ambient Traffic Pollution in Toronto, Canada", *Strategies for Clean Air and Health Conference.* Rome, Italy. http://www.irr-neram.ca/rome/Proceedings/Jerrett.pdf.

Jobling, M.A. (2001). "In the Name of the Father: Surnames and Genetics", *Trends in Genetics*, 17(6): 353–357.

Johnson, I. (2004). "Putting Time on the Map: Using TimeMap for Map Animation and Web Delivery", *GeoInformatics,* 7(5): 26–29.

Jones, C.B., Abdelmoty, A.I., Finch, D., Fu, G. and Vaid, S. (2004). "The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing", *Geographic Information Science: Third International Conference* (GIScience-2004). Adelphi, PA, 3234: 125–139.

Jones, C.B., Alani, H. and Tudhope, D. (2001). "Geographical Information Retrieval with Ontologies of Place", *International Conference on Spatial Information Theory: Foundations of Geographic Information Science (LNCS, Vol. 2205)*. Ed. D.R. Montello. Berlin: Springer, 322–335.

JRC (2005). "Population Density Disaggregated with CLC2000, Version 3", Joint Research Centre. Ispra, Italy. http://dataservice.eea.europa.eu/dataservice/metadetails.asp?id=830/.

Julien, C. and Roman, G.C. (2002). "Egocentric Context-Aware Programming in Ad-Hoc Mobile Environments", *Symposium on Foundations of Software Engineering*. Charleston, SC.

Jurafski, D. and Martin, J.H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Upper Saddle River, NJ: Prentice Hall.

Kadobayashi, R. and Tanaka, K. (2005). "3D Viewpoint-based Photo Search and Information Browsing", *28th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR-2005)*. New York: ACM Press, 621–622.

Kashyap, V. and Sheth, A. (1996). "Schematic and Semantic Similarities between Database Objects: A Context-based Approach," *Very Large Data Bases (VLDB) Journal*, 5(4): 276–304.

Keller, R.M., Berrios, D.C., Carvalho, R.E., Hall, D.R., Rich, S.J., Sturken, I.B., Swanson, K.J. and Wolfe, S.R. (2004). "SemanticOrganizer: A Customizable Semantic Repository for Distributed NASA Project Teams", *3rd International - Semantic Web Conference* (ISWC-2004). Hiroshima, Japan: Springer, 767–781.

Kelmelis, J., DeMulder, M., Ogrosky, C., Van Driel, N. and Ryan, B. (2003). "The National Map from Geography to Mapping and Back Again", *Photogrammetric Engineering and Remote Sensing,* 69(10): 1109–1111.

Kendall, J.E. (2005). "Satellite Mapping and Its Potential in Ecommerce: Why We Need Directions to Follow Our New Maps", *Decision Line,* 36(5): 11–14.

Kienreich, W., Granitzer, M. and Lux, M. (2006). "Geospatial Anchoring of Encyclopedia Articles", *10th International Conference on Information Visualisation (iV-06)*. London, UK: In print.

Kingston, R. (2002). "Web-based PPGIS in the United Kingdom", *Community Participation and Geographic Information Systems.* Eds. W.J. Craig et al. New York: Taylor and Francis, 101–112.

Kishor, P. and Ventura, S. (2006). "What Can GIS Learn from FLOSS?" *9th International Conference of the Global Spatial Data Infrastructure* (*GSDI-9-2006)*. Santiago, Chile. http://www.gsdi9.cl/english/papers/TS50.4paper.pdf.

Kleppin, L. (2006). Integration der Ergebnisse des bundesweiten UNECE Moos-Monitorings in ein Internet-gestütztes Geo-Informationssystem (WebGIS) und deren statistische Analyse mittels Chi-square Automatic Interaction Detection (CHAID)", *Diploma Thesis*. University of Vechta.

Klien, E., Einspanier, U., Lutz, M. and Hübner, S. (2004). "An Architecture for Ontology-based Discovery and Retrieval of Geographic Information", *Semantic Web Services and Dynamic Networks Workshop (SWSDN-2004)*. Ulm, Germany, 574–578.

Kolas, D. and Kammersell, W. (2005). "*Semwebcentral: Project Info – GeoSwrl*". http://projects.semwebcentral.org/projects/geoswrl/.

Kolas, D., Hebeler, J. and Dean, M. (2005). "Geospatial Semantic Web: Architecture of Ontologies", *First International Conference on Geospatial Semantics* (*GeoS-2005), Mexico City (LNCS, Vol. 3799)*. Springer, 183–194.

Konecný, M. and Bandrova, T. (2006). "Proposal for a Standard in Cartographic Visualization of Natural Risks and Disasters", *Joint Symposium of Seoul Metropolitan Fora and Second International Workshop on Ubiquitous, Pervasive and Internet Mapping*. Seoul, Korea, 165–173.

Konecný, M., Kolár, M., Brázdilová, J., Kolejka, J., Michálek, J., Talhofer, V. and Svancara, J. (2005). "Dynamická Geovizualizace v Krizovém Managementu (Dynamic Geovisualization in Crises Management)", *Project No. MSM 0021622418 (Supported by the Ministry of Education, Youth and Sports of the Czech Republic)*. http://geokrima.geogr.muni.cz/index.html.

Krygier, J.B. (1999). "World Wide Web Mapping and GIS: An Application for Public Participation", *Cartographic Perspectives,* 33: 66–67.

Kubícek, P. and Stanek, K. (2006). "Dynamic Visualization in Emergency Management", *1st International Conference on Cartography and GIS*. Ed. T. Bandrova. Borovets, Bulgaria, 25–28.

Kuhn, W. (2002). "Modeling the Semantics of Geographic Categories through Conceptual Integration", *Geographic Information Science-Second International Conference*, GIScience 2002 (LNCS, Vol. 2478). Boulder, CO: Springer.

Kuhn, W. (2005). "Geospatial Semantics: Why, of What, and How?" *Journal of Data Semantics,* 3: 1–24.

Kunreuther, H. and Grossi, P. (2005). *Catastrophe Modeling: A New Approach to Managing Risk*. New York: Springer.

Kutz, D. and Herring, S.C. (2005). "Micro-Longitudinal Analysis of Web News Updates", *38th Hawaii International Conference on System Sciences (HICSS-38-2005)*. Hawaii: IEEE Press.

Lake, R., Burggraf, D., Trninic, M. and Rae, L. (2004). *Geography Mark-Up Language: Foundation for the Geo-Web*. New York: Wiley.

Larson, R.R. (1996). "Geographic Information Retrieval and Spatial Browsing", *GIS and Libraries: Patrons, Maps and Spatial Information*. Eds. L. Smith and M. Gluck. Urbana-Champaign, IL: University of Illinois Press, 81–124.

Leidner, J.L. (2004). "Towards a Reference Corpus for Automatic Toponym Resolution Evaluation", *27th Annual International ACM SIGIR Conference (SIGIR-2004), Workshop on Geographic Information Retrieval.* http://www.geo.unizh.ch/~rsp/gir/abstracts/leidner.pdf.

Lemmens, R. (2006). "Semantic Interoperability of Distributed Geo-Services", *PhD Thesis*, Enschede, The Netherlands.

Lemmens, R., Wytzisk A., de By, R., Granell, C., Gould, M. and van Oosterom, P. (2006). "Integrating Semantic and Syntactic Descriptions for Chaining Geographic Services", *IEEE Internet Computing,* 10(5): 42–52.

Lesbegueries, J., Gaio, M., Loustau, P. and Sallaberry, C. (2006). "Geographical Information Access for Non-Structured Data", *ACM Symposium on Applied Computing (SAC-2006)*. Dijon, France, 83–89.

Lesbegueries, J., Sallaberry, C. and Gaio, M. (2006). "Associating Spatial Patterns to Text-Units for Summarizing Geographic Information", *29th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval – GIR (Geographic Information Retrieval) Workshop*. Seattle, WA. http://www.geo.unizh.ch/~rsp/gir06/papers/individual/lesbegueries.pdf.

Letzel, S. and Gacki, R. (2001). *PHP und MySQL*. München, Germany: Markt und Technik Verlag.

Levy, S. (2004). "Making the Ultimate Map", *Newsweek,* 143(23): 56–58.

Li, B. and Fonseca, F.T. (2006). "TDD – a Comprehensive Model for Qualitative Spatial Similarity Assessment", *Spatial Cognition and Computation*, 6(1): 31–62.

Li, H., Srihari, R.K., Niu, C. and Li ,W. (2003). "InfoXtract Location Normalization: A Hybrid Approach to Geographic References in Information Extraction", *Workshop on Analysis of Geographic References – Volume 1, Human Language Technology Conference* (*HLT-NAACL-2003)*. Morristown, NJ: Association for Computational Linguistics, 39–44.

Li, X., Morie, P. and Roth, D. (2005). "Semantic Integration in Text: From Ambiguous Names to Identifiable Entities", *AI Magazine*: *Special Issue on Semantic Integration*, 45–68.

Liebhold, M. (2004). Infrastructure for the New Geography. Menlo Park, CA: Institute for the Future. http://www.iftf.org/docs/SR-869_Infra_New_Geog_Intro.pdf.

Liebig, W. and Mummenthey, R.-D. (2005). *ArcGIS-ArcView9 – ArcGIS Grundlagen, Bd. 1.* Norden, Germany: Points Verlag.

Lloyd, C.D. and Atkinson, P.M. (2004). "Increased Accuracy of Geostatistical Prediction of Nitrogen Dioxide in the United Kingdom with Secondary Data", *International Journal of Applied Earth Observation and Geoinformation,* 5: 293–305.

Loibl, W., Züger, J. and Kutschera, P. (2000). "Verbesserung des Modells zur Generierung von stündlichen Karten der Ozonkonzentration für Österreich", *Seibersdorf Research Report*, OEFZS-S-0063, BV, 64.

Loustau, P. (2005). "Traitements Sémantiques de Documents dans Leur Composante Spatiale", *Master's thesis.* Université de Pau et des Pays de l'Adour France.

Lowe, A. (2003). "The Federal Emergency Management Agency's Multi-Hazard Flood Map Modernization and the National Map", *Photogrammetric Engineering and Remote Sensing*, 69(10): 1133–1135.

Luck, M., McBurney, P., Shehory, O. and Willmott, S. (2005). "Agent Technology Roadmap: A Roadmap for Agent-based Computing". http://www.agentlink.org/roadmap/index.html.

Lutz, C. and Wolter, F. (2006). "Modal Logics of Topological Relations", *Logical Methods in Computer Science*, 2(2): 1–41.

Lutz, M. and Klien, E. (2006). "Ontology-based Retrieval of Geographic Information", *International Journal of Geographical Information Science,* 20(3): 233–260.

MacEachren, A.M. (1995). *How Maps Work: Representation, Visualisation and Design.* New York: The Guilford Press.

MacEachren, A.M., Cai, G., McNeese, M., Sharma, R. and Fuhrmann, S. (2006). "GeoCollaborative Crisis Management: Designing Technologies to Meet Real-World Needs", *7th Annual National Conference on Digital Government Research*: *Integrating Information Technology and Social Science Research for Effective Government*. San Diego, CA, 71–72.

Maguire, D. (2006). "Geographic Earth Explorers: A New Software Paradigm for Visualizing and Analyzing Geography?" *2006 Annual Meeting of the Association of American Geographers*. Chicago, IL.

Maguire, D. (2006). "GeoWeb 2.0". http://gismatters.blogspot.com/2006/06/-geoweb-20.html.

Mamei, M., Quaglieri, R. and Zambonelli, F. (2006). "Making Tuple Spaces Physical with RFID Tags", *ACM Symposium on Applied Computing*. Dijon, France, 434–439.

Manning, C.D. and Schütze H. (1999). *Foundations of Statistical Natural Language Processing.* 1st Edition. Cambridge, MA: MIT Press.

Marcellin, M.W. Gormish, M.J., Bilgin, A. and Boliek, M.P. (2000). "An Overview of JPEG-2000", *IEEE Data Compression Conference (DCC-2000).* Salt Lake City, UT: IEEE. http://rii.ricoh.com/~gormish/pdf/dcc2000_jpeg2000_note.pdf.

Mark, D.M. (1989). "Cognitive Image-Schemata for Geographic Information: Relations to User Views and GIS Interfaces". *Annual Conference and Exposition (GIS/LIS-1989)*. Orlando, FL, 551–560.

Mark, D.M. (1993). "On the Ethics of Representation, or Who's World Is It, Anyway?" *NCGIA Geographic Information and Society Workshop,* Friday Harbor, WA.

Markowetz, A., Chen, Y.-Y., Suel, T., Long, X. and Seeger, B. (2005). "Design and Implementation of a Geographic Search Engine", *8th International Workshop on the Web and Databases* (WebDB-2005). Baltimore, MD. http://cis.poly.edu/suel/papers/geo.pdf.

Marquesuzaà, C., Etcheverry, P. and Lesbegueries, J. (2005). "Exploiting Geospatial Markers to Explore and Resocialize Localized Documents", *First International Conference GeoSpatial Semantics.* Mexico City: Springer, 153–165.

Martin, D., Burstein, M., Hobbs, J., Lassila, O., McDermott, D., McIlraith, S., Narayanan, S., Paolucci, M., Parsia, B., Payne, T., Sirin, E., Srinivasan, N. and Sycara, K. (2004). "OWL-S: Semantic Markup for Web Services", *W3C Member Submission*. http://www.w3.org/Submission/OWL-S/.

Mason, B.C. and Dragicevic, S. (2006). "Web GIS and Knowledge Management Systems: An Integrated Design for Collaborative Community Planning", *Collaborative Geographic Information Systems*. Eds. S. Balram and S. Dragicevic. Hershey, PA: Idea Group, 263–284.

Maynard, D., Tablan, V., Bontcheva, K., Cunningham, H. and Wilks, Y. (2003). "MUSE: A Multi-Source Entity Recognition System". Netherlands: Kluwer Academic Publishers. http://gate.ac.uk/sale/muse/muse.pdf.

McCurdy, N.J. and Griswold, W.G. (2005). "A Systems Architecture for Ubiquitous Video", *3rd International Conference on Mobile Systems, Applications, and Services (MobiSys-2005)*. New York: ACM Press, 1–14.

McCurley, K.S. (2001). "Geospatial Mapping and Navigation of the Web", *10th International World Wide Web Conference (WWW-2001)*. Hong Kong: ACM Press, 221–229.

McLuhan, M. (1964). *Understanding Media: The Extension of Man*. London: Routledge and Kegan Paul.

Mengual, P. (2005). "Contribution à la Caractérisation de la Vulnérabilité des PME-PMI aux Inondations", *Thesis Report*. Nice, France: Sophia-Antipolis University.

Mennis, J.L., Peuquet, D.J. and Qian, L. (2000). "A Conceptual Framework for Incorporating Cognitive Principles into Geographical Database Representation", *International Journal of Geographical Information Science*, 14(6): 501–520.

Meyer, V. and Messner, F. (2005). "Flood Damage, Vulnerability and Risk Perception – Challenges for Flood Damage Research", *Flood Risk Management – Hazards, Vulnerability and Mitigation Measures, NATO Science Series.* Eds. J. Schanze et al. Berlin: Springer.

Michalowski, M., Ambite, J., Thakkar, S., Tuchinda, R., Knoblock, C. and Minton, S. (2004). "Retrieving and Semantically Integrating Heterogeneous Data from the Web", *IEEE Intelligent Systems*, 19: 72–79.

Mikheev, A., Grover, C. and Moens, M. (1998). "Description of the LTG System Used for MUC-7", *7th Message Understanding Conference*. Fairfax, VA. http://www-nlpir.nist.gov/related_projects/muc/.

Mills, E. (2005). "Mapping a Revolution with Mashups", *CNET News*. http://news.-com.com/mapping+a+revolution+with+mashups/2009-1025_3-5944608.html.

Mitchell, T. (2005). *Web Mapping Illustrated*. Sebastopol, CA: O´Reilly Media.

Moodley, D. and Kinyua, J.D.M. (2006). "Towards a Multi-Agent Infrastructure for Distributed Internet Applications", *8th Annual Conference on World Wide Web Applications*. Bloemfontein, South Africa. http://www.zaw3.co.za.

Moodley, D., Terhorst, A., Simonis, I., McFerren, G. and van den Bergh, F. (2006). "Using the Sensor Web to Detect and Monitor the Spread of Wild Fires", *2nd International Symposium on Geographic Information for Disaster Management* (Gi4DM-2006). Goa, India.

Moore, C. and Newman, M.E.J. (2000). "Epidemics and Percolation in Small-World Networks", *Physical Review E*, 61: 5678–5682.

Morimoto, Y., Aono, M., Houle, M.E. and McCurley, K.S. (2003). "Extracting Spatial Knowledge from the Web", *Symposium on Applications and the Internet (SAINT-2003)*. Orlando, FL: IEEE Computer Society, 326–333.

Morita, T. (2004). "Ubiquitous Mapping in Tokyo", *International Joint Workshop on Ubiquitous, Pervasive and Internet Mapping (*UPIMap-2004). Tokyo, Japan. http://www.ubimap.net/upimap2004/html/papers/UPIMap04-A-01-Morita.pdf.

Motta, E. (1999). *Reusable Components for Knowledge Modelling*. Amsterdam: IOS Press.

Motta, E., Shum, S.B. and Domingue, J. (2000). "Ontology-driven Document Enrichment: Principles, Tools and Applications", *International Journal of Human-Computer Studies,* 52(6): 1071–1109.

Müller, M., Vorogushyn, S., Maier, P., Thieken, A., Petrow, T., Kron, A., Büchele, B. and Wächter, J. (2006). "CEDIM Risk Explorer – a Map Server Solution in the Project Risk Map Germany", *Natural Hazards Earth System Science*, 6: 711–720.

Muller, P. (2002). "Topological Spatio-Temporal Reasoning and Representation", *Computational Intelligence*, 18(3): 420–450.

Na, A. and Priest, M. (2006). "Sensor Observation Service", *Document 05-088r1*, Wayland, MA. http://portal.opengeospatial.org/files/?artifact id=12846.

Naaman, M., Paepcke, A. and Garcia-Molina, H. (2003). "From Where to What: Metadata Sharing for Digital Photographs with Geographic Coordinates", *Lecture Notes in Computer Science,* Berlin, Germany: Springer, 2888: 196–217.

Neches, R. Yao, K.-T., Bugacov, A., Kumar, V. and Eleish, R. (2001). "GeoWorlds: Integrating GIS and Digital Libraries for Situation Understanding and Management", *New Review of Hypermedia and Multimedia,* 7: 127–152.

Nejdl, W. and Wolf , B. (2002). "EDUTELLA: A P2P Networking Infrastructure Based on RDF", *Proceedings of the World Wide Web Conference*. Honolulu, HI. http://citeseer.csail.mit.edu/boris01edutella.html.

Nejdl, W., Wolpers, M., Siberski, W., Schmitz, C., Schlosser, M., Bruckhorst , I. and Löser, A. (2003) "SuperPeer-based Routing and Clustering Strategies for RDF-based Peer-to-Peer Networks", *12th World Wide Web Conference*. Budapest, Hungary. http://citeseer.ist.psu.edu/article/nejdl03superpeerbased.html.

Neteler, M. and Mitasova, H. (2004). *Open Source GIS: A GRASS GIS Approach*. 2nd Edition. Boston, MA: Springer.

New, M., Lister, D., Hulme, M. and Makin, I. (2002). "A High-resolution Data Set of Surface Climate over Global Land Areas", *Climate Research,* 21: 1–25.

Newitz, A. (2006). "Map Mashups Get Personal", *Wired*. http://www.wired.com/news/technology/1,70419-0.html.

Newman, M.E.J. (2003). "The Structure and Function of Complex Networks", *SIAM Review,* 45: 167–256.

NGA (2006). *GEOnet Names Server (GNS),* National Geospatial-Intelligence Agency. http://earth-info.nga.mil/gns/html/index.html.

Niles, I. and Pease, A. (2001). "Towards a Standard Upper Ontology", *2nd International Conference on Formal Ontology in Information Systems* (*FOIS-2001).* Eds. C. Welty and B. Smith. Ogunquit, ME. http://citeseer.ist.psu.edu/-niles01towards.html.

Nodenot, T., Loustau, P., Gaio, M., Sallaberry, C. and Lopisteguy, P. (2006). "From Electronic Documents to Problem-based Learning Environments: An Ongoing Challenge for Educational Modelling Languages", *Conference on Information Technology Based Higher Education & Training (ITHET-2006). Sydney, Australia.*

NRC (1993). *Toward a Coordinated Spatial Data Infrastructure for the Nation.* National Research Council – Mapping Science Committee. Washington, DC: National Academy Press.

Nussbaum, R. (2000). "Pourquoi une Mission Risques Naturels?", *Risques Revue,* 42. http://www.ffsa.fr/webffsa/risques.nsf/.

OGC (1999). *Simple Features Specification for SQL*. http://portal.opengeospatial.org/ files/?artifact_id=829.

OGC (2003). *OGC Reference Model*. Ed. G. Percivall. http://portal.opengeospatial.org/files/?artifact_id=3836.

OGC (2006a). *OpenGIS® Catalog Services — ebRIM (ISO/TS 15000-3). Profile of CSW*. Ed. R. Martell. http://www.opengis.org/docs/05-025r3.pdf.

OGC (2006b). *GML in JPEG 2000 for Geographic Imagery (GMLJP2) Encoding Specification*. http://www.opengeospatial.org/standards/gmljp2.

OGC (2006c). *OGC Abstract Specifications.* http://www.opengis.org/techno/abstract.htm.

OGC (2006d). *OpenGIS Registered Products.* http://www.opengeospatial.org/resources/?page=products.

Ogren, P., Fiorelli, E. and Leonard, N.E. (2004). "Cooperative Control of Mobile Sensor Networks: Adaptive Gradient Climbing in a Distributed Environment." *IEEE Transactions on Automatic Control*, 49: 1292–1302.

Olson, G.M., Atkins, D.E., Clauer, R., Finholt, T.A., Jahanian, F., Killeen, T.L., Prakash, A. and Weymouth, T. (1998). "The Upper Atmospheric Research Collaboratory." *Interactions*, 5: 48–55.

Oracle (2005). *Oracle® Spatial User's Guide and Reference*. http://www.oracle.com/technology/documentation/spatial.html.

O'Reilly, T. (2005). "What Is Web 2.0 – Design Patterns and Business Models for the Next Generation of Software". San Francisco, CA: O'Reilly. http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html.

Parkinson, C.L. (2003). "Aqua: An Earth-Observing Satellite Mission to Examine Water and Other Climate Variables", *IEEE Transactions on Geoscience and Remote Sensing*, 41(2): 173–183.

Parr, T.W., Ferretti, M., Simpson, I.C., Forsius, M. and Kóvacs-Láng, E. (2002). "Towards a Long-Term Integrated Monitoring Programme in Europe. Network Design in Theory and Practise", *Environmental Monitoring and Assessment,* 78: 253–290.

Pavlik, J.V. (1998). *New Media Technology – Cultural and Commercial Perspectives*. Needham Heights, MA: Allyn and Bacon.

Peng, Z.-R. and Tsou, M.-H. (2003). *Internet GIS: Distributed Geographic Information Services for the Internet and Wireless Networks*. Hoboken, NJ: Wiley.

Perry, M., Hakimpour, F. and Sheth, A. (2006). "Analyzing Theme, Space, and Time: An Ontology-based Approach", *14th International ACM Symposium on Advances in Geographic Information Systems (GIS-2006)*. Arlington, VA, 147–154.

Pesch, R. (2003). "Geostatistische und multivariat-statistische Analyse der Ergebnisse des Moos-Monitorings 1990, 1995 und 2000 zur Ableitung von Indikatoren für atmosphärische Metalleinträge in Deutschland", *PhD Dissertation*. University of Vechta.

Pesch, R. and Schröder, W. (2005). "Integrative Exposure Assessment through Classification and Regression Trees on Bioaccumulation of Metals, Related Sampling Site Characteristics and Ecoregions". *Ecological Informatics,* 1(1): 55–65.

Pesch, R. and Schröder, W. (2006). "Spatiotemporal Variability of Metal Accumulation in Mosses. Analysis of Measurement Data and Metadata by Statistics and GIS", *Nova Hedwegia,* 82: 447–466.

Peuquet, D.J. (2002). *Representations of Space and Time*. New York: Guilford.

Pickles, J. (1991). "Geography, GIS and the Surveillant Society", *Papers and Proceedings of the Applied Geography Conferences*. Eds. J.W. Frazier et al. New York: State University NY Press, 14: 80–91.

Pickles, J. (1997). "Tool or Science? GIS, Technoscience, and the Theoretical Turn", *Annals of the Association of American Geographers,* 87(2): 363–372.

Pickles, J. (1999). "Arguments, Debates, and Dialogues: The GIS-Social Theory Debate and the Concern for Alternatives", *Geographic Information Systems, Principles and Technical Issues.* Eds. P.A. Longley et al. New York: Wiley, 1: 49–60.

Pietriga, E. (2002). "Isaviz: A Visual Environment for Browsing and Authoring RDF Models ", *11th International World Wide Web Conference (WWW-2002)*. Honolulu, HI.

Prud'hommeaux, E. and Seaborne, A. (2006). "SPARQL Query Language for RDF", *W3C Candidate Recommendation*. http://www.w3.org/TR/rdf-sparql-query/.

Quan, D. and Karger, D.R. (2004)."How to Make a Semantic Web Browser", *13th International World Wide Web Conference (WWW-2004)*. New York. http://citeseer.ist.psu.edu/quan04how.html.

Quan, D., Huynh, D. and Karger, D.R. (2003). "Haystack: A Platform for Authoring End User Semantic Web Applications", *International Semantic Web Conference* (ISWC-2003). Sanible Island, FL (LNCS, Vol. 2870). Eds. D. Fensel et al. Berlin: Springer, 738–753.

Ramsey, P. (2004). "A Survey of OGC Deployment". http://www.digitalearth.org/story/2004/12/1/15658/1000.

Randell, D.A., Cui, Z. and Cohn, A.G. (1992). "A Spatial Logic Based on Regions and Connections", *3rd International Conference on Knowledge Representation and Reasoning*. San Mateo, CA: Morgan Kaufmann, 165–176.

Raskin, R. (2006). "Guide to SWEET Ontologies for Earth System Science", *Technical Report*. http://sweet.jpl.nasa.gov/guide.doc.

Rauch, E., Bukatin, M. and Baker, K. (2003). "A Confidence-based Framework for Disambiguating Geographic Terms", *Workshop on the Analysis of Geographic References, Human Language Technology Conference (NAACL/HTL-2003)*. Edmonton, Canada, 50–54.

Redner, S. (1998). "How Popular Is Your Paper? An Empirical Study of the Citation Distribution", *The European Physical Journal B,* 4: 131–134.

Rodden, K. and Wood, K. (2003). "How Do People Manage Their Digital Photographs?" *Conference on Human Factors in Computing Systems (SIGCHI-2003)*. New York: ACM Press, 24–26.

Rodríguez, A.M. and Egenhofer, M.J. (2004). "Comparing Geospatial Entity Classes: An Asymmetric and Context-Dependent Similarity Measure", *International Journal of Geographical Information Science,* 18(3): 229–256.

Roman, D., Keller, U., Lausen, H., Lara, R., de Bruijn, J., Stollberg, M., Polleres, A., Feier, C., Bussler, C. and Fensel, D. (2005). "Web Service Modeling Ontology". *Applied Ontology*, 1(1): 77–106.

Roman, D., Lausen, H. and Keller, U. (2004). "Web Service Modeling Ontology Standard". *Technical Report, WSMO Working Draft*. http://www.wsmo.org/2004/d2/v0.2/20040306/.

Roush, W. (2005). "Killer Maps", *Technology Review,* 108(10): 54–60.

Rüger, S. (2005). "Putting the User in the Loop: Visual Resource Discovery", *3rd International Workshop on Adaptive Multimedia Retrieval (AMR-2005)*. Glasgow, UK. http://mmir.doc.ic.ac.uk/www-pub/amr-2005.pdf.

Rühling, A. (1994). "Atmospheric Heavy Metal Deposition in Europe – Estimations Based on Moss Analysis", *Nordic Council of Ministers,* 9. http://www.norden.org/.

Rühling, A. and Steinnes, E. (1998). "Atmospheric Heavy Metal Deposition in Europe 1995–1996", *Nordic Council of Ministers,* 15. http://www.norden.org/.

Runciman, B. (2006). "Interview with Tim Berners-Lee", *ITNOW*, 48(2): 18–21.

Sabidussi, G. (1966). "The Centrality Index of a Graph", *Psychometrika*, 31: 581–603.

Sabol, V., Kienreich, W., Granitzer, M., Becker, J., Tochtermann, K. and Andrews, K. (2002). "Applications of a Lightweight, Web-based Retrieval, Clustering, and Visualisation Framework", *4th International Conference on Practical Aspects of Knowledge Management (LNCS, Vol. 2569)*. Eds. D. Karagiannis and U. Reimer. Berlin: Springer, 359–368.

Sallaberry, C., Etcheverry, P. and Marquesuzaà, C. (2006). "Information Retrieval and Visualization Based on Documents' Geospatial Semantics", *4th IEEE International Conference on Information Technology: Research and Education (ITRE-2006)*. Tel Aviv, Israel.

Sallaberry, C., Marquesuzaà, C. and Etcheverry, P. (2006). "Spatial Information Management within Digital Libraries", *1st IEEE International Conference on Digital Information Management (ICDIM-2006)*. Bangalore, India.

Sanderson, M. and Kohler, J. (2004). "Analyzing Geographic Queries". *Workshop on Geographic Information Retrieval (SIGIR-2004)*. Sheffield, UK. http:// citeseer.ist.psu.edu/sanderson04analyzing.html.

Sarigiannis, D., Sifakis, N., Soulakellis, N., Schafer, K. and Tombrou, M. (2003). "A New Approach to Environmental Data Fusion for Integrated Assessment of Particulate Matter Loading and Its Effect on Health in the Urban Environment". *IEEE International Geoscience and Remote Sensing Symposium (IGARSS-2003).* Toulouse, France, 7: 4579–4581.

Satoh, I. (2005). "A Location Model for Pervasive Computing Environments", *International Conference on Pervasive Computing and Communications (PerCom-2005)*. Kauai, HI, 215–224.

Schade, S. (2005). "Sensors on the Way to Semantic Interoperability", *GI-Days 2005: Geo-Sensor Networks – from Science to Practical Applications.* ifgi-Prints, Vol. 23. Münster: University of Münster.

Scharl, A. (2000). *Evolutionary Web Development*. London: Springer. http://webdev.wu-wien.ac.at/.

Scharl, A. (2001). "Explanation and Exploration: Visualizing the Topology of Web Information Systems", *International Journal of Human-Computer Studies,* 55(3): 239–258.

Scharl, A. (2004). "Web Coverage of Renewable Energy", *Environmental Online Communication.* Ed. A. Scharl. London: Springer. 25-34.

Scharl, A., Weichselbraun, A. and Liu, W. (2005). "An Ontology-based Architecture for Tracking Information across Interactive Electronic Environments", *39th Hawaii International Conference on System Sciences.* Kauai, HI: IEEE Press.

Scheibner, H., Appelrath, H.J., Jobmann, K. and Reimers, U. (2006). *Niccimon – Das Niedersächsische Kompetenzzentrum Informationssysteme für die Mobile Nutzung. Bericht und Ausblick.* Aachen: Shaker.

Schilit, B., Adams, N. and Want, R. (1994). "Context-Aware Computing Applications", *Workshop on Mobile Computing Systems and Applications*. English Lake District, UK.

Schmid, H. (1994). "Probabilistic Part-of-Speech Tagging Using Decision Trees", *Conference on New Methods in Language Processing*. Manchester, UK. http://citeseer.comp.nus.edu.sg/schmid94probabilistic.html.

Schmidt, G. (2002). "Eine multivariat-statistisch abgeleitete ökologische Raumgliederung für Deutschland". PhD thesis. Berlin: dissertation.de.

Schröder, W. and Pesch, R. (2004a). "Spatial Analysis and Indicator Building for Metal Accumulation in Mosses". *Environmental Monitoring and Assessment*, 98: 131–155.

Schröder, W. and Pesch, R. (2004b). "Integrative Monitoring Analysis Aiming at the Detection of Spatial and Temporal Trends of Metal Accumulation in Mosses". *Journal of Atmospheric Chemistry*, 49: 23–38.

Schröder, W., Anhelm, P., Bau, H., Matter, Y., Mitze, R., Mohr, K., Peichl, L., Peiter, A., Peronne, T., Pesch, R., Roostai, H., Roostai, Z., Schmidt, G. and Siewers, U. (2002). "Untersuchungen von Schadstoffeinträgen anhand von Bioindikatoren", *Abschlussbericht FuE-Vorhaben 200 64 218*. Berlin: Umweltbundesamt.

Schroeder, P.C. (1996). "Public Participation GIS: Local and Regional Potential of Spatial Technologies", *7th Annual Conference at the Atlantic Institute*. Orono, ME.

Schutzberg, A. (2005). "Geographic (and Other Types) of Metadata in the Newsroom", *Directions Magazine*. August 08, 2005. http://www.directionsmag.com/-article.php?article_id=931.

Schuyler, E. (2005). *Mapping Hacks, Tips and Tools for Electronic Cartography*. Sebastopol, CA: O´Reilly.

Schwering, A. (2005). "Hybrid Model for Semantic Similarity Measurement", *4th International Conference on Ontologies, Databases, and Applications of Semantics (*ODBASE-2005). Agia Napa, Cyprus: Springer, 1449–1465.

Shadbolt, N., Berners-Lee, T. and Hall, W. (2006). "The Semantic Web Revisited", *IEEE Intelligent Systems*, 21(3): 96–101.

Shahabi, C., Kolahdouzan, M.R., Thakkar, S., Ambite, J.L. and Knoblock, C.A. (2001). "Efficiently Querying Moving Objects with Pre-defined Paths in a Distributed Environment", *9th ACM International Symposium on Advances in Geographic Information Systems*. Atlanta, GA: ACM Press, 34–40.

Shapiro, C. and Varian, H.R. (1999). "The Art of Standards War", *California Management Review,* 41(2): 8–32.

Shariff, A.R., Egenhofer, M.J. and Mark, D.M. (1998). "Natural Language Spatial Relations between Linear and Areal Objects: The Topology and Metric of English Language Terms", *International Journal of Geographical Information Science*, 12(3): 215–246.

Sheth, A. (1999). "Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics", *Interoperating Geographic Information Systems*. Eds. M.F. Goodchild et al. Norwell, MA: Kluwer Academic Publishers, 5–30.

Shiffer, M.J. (1999). "Managing Public Discourse: Towards the Augmentation of GIS with Multimedia", *Geographic Information Systems, Management Issues and Applications*. Eds. P. A. Longley et al. New York: Wiley, 2: 723–732.

Shneiderman, B. and Plaisant, C. (2004). *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Boston, MA: Addison-Wesley.

Sieber, R. (2006). "Public Participation Geographic Information Systems: A Literature Review and Framework", *Annals of the Association of American Geographers*, 96(3): 491–507.

Sierhuis, M., Clancey, W.J. and Sims, M.H. (2002). "Multiagent Modeling and Simulation in Human-Robot Mission Operations Work System Design", *35th Annual Hawaii International Conference on System Sciences*. Ed. R.H. Sprague. Kona, HI: IEEE Computer Society Press, 191–200.

Siewers, U. and Herpin, U. (1998). "Schwermetalleinträge in Deutschland. Moos-Monitoring 1995", *Geologisches Jahrbuch*, Sonderhefte, Heft SD 2. Stuttgart: Bornträger.

Siewers, U., Herpin, U. and Strassburger, S. (2000). "Schwermetalleinträge in Deutschland. Moos-Monitoring 1995", Teil 2, *Geologisches Jahrbuch*, Sonderhefte, Heft SD 3. Stuttgart: Bornträger.

Simonis, I. (2005). "Sensor Planning Service". *Document 05-089r1*. Wayland, MA. http://portal.opengeospatial.org/files/?artifact id=12971.

Simonis, I. (2006). "Sensor Alert Service", *Document 06-028*. Wayland, MA. http://portal.opengeospatial.org/files/?artifact id=13921.

Simonis, I. and Wytzisk, A. (2003). "Web Notification Service", *Document 03-008r2*. Wayland, MA. http://portal.opengeospatial.org/files/?artifact id=1367.

Sirin, E. (2004). *OWL-S API*. http://www.mindswap.org/2004/owl-s/api/.

Smart, P., Abdelmoty, A.I. and Jones, C.A. (2006). "Visual Editor for Validating Geo-Ontologies in OWL", *GIS Research UK Conference (GISRUK-2006)*. Nottingham, UK.

Smith, B. and Mark, D.M. (1999). "Ontology with Human Subjects Testing: An Empirical Investigation of Geographic Categories", *American Journal of Economics and Sociology*, 58: 245–272.

Smith, D.A. and Crane, G. (2001). "Disambiguating Geographic Names in a Historical Digital Library", *5th European Conference on Research and Advanced Technology for Digital Libraries*. Darmstadt, Germany, 127–136.

Smith, D.A. and Mann, G.S. (2003). "Bootstrapping Toponym Classifiers", *HTL/NAACL Workshop on the Analysis of Geographic References (HTL/NAACL-2003)*. Edmonton, Canada, 45–49.

Smith, M.K., Welty, C. and McGuinness, D. (2004). "Web Ontology Language (OWL) Guide Version 1.0", *World Wide Web Consortium.* http://www.w3.org/TR/2004/REC-owl-guide-20040210/.

Snavely, N., Seitz, S.M. and Szeliski, R. (2006). "Photo Tourism: Exploring Photo Collections in 3D", *ACM Transactions on Graphics (SIGGRAPH-2006)*. New York: ACM Press, 835–846.

Spärck Jones, K. (1972). "A Statistical Interpretation of Term Specificity and Its Application in Retrieval." *Journal of Documentation,* 28: 11–21 and 60: 493–502.

Spohrer, J. (1999). "Information in Places", *IBM Systems Journal,* 38(4): 602–628.

Stapleton, C. and Hughes, C.E. (2006). "Believing Is Seeing: Cultivating Radical Media Innovations", *IEEE Computer Graphics and Applications,* 26(1): 88–93.

Stedman, J.R., Bush, T.J., Murrels, T.P. and King, K. (2001). "Baseline PM_{10} and NO_x Projections for PM_{10} Objective Analysis", *AEA Technology*, National Environmental Technology Centre. Report AEAT/ENV/R/0726. http://www.aeat.co.uk/netcen/airqual/reports/naqs2001/aeat-env-r-0726.pdf.

Stephenson, N. (1992). *Snow Crash*. New York: Bantam Books.

Stonebraker, M. (1996). *Object-Relational DBMS: The Next Great Wave*. San Francisco, CA: Morgan-Kaufmann.

Strogatz, S.H. (2001). "Exploring Complex Networks", *Nature,* 410: 268–276.

Swann, G.M.P. (2002). "The Functional Form of Network Effects", *Information Economics and Policy,* 14(3): 417–429.

Tan, P.-N., Steinbach, M. and Kumar, V. (2005). *Introduction to Data Mining*. Boston, MA: Addison-Wesley.

Tang, T.T., Hawking, D., Craswell, N. and Sankaranarayana, R.S. (2004). "Focused Crawling in Depression Portal Search: A Feasibility Study", *9th Australasian Document Computing Symposium (ADCS 2004)*. Eds. P. Bruza et al. Melbourne: University of Melbourne, 2–9.

Tate, A., Buckingham Shum, S., Dalton, J., Mancini, C. and Selvin, A. (2006). "Co-OPR: Design and Evaluation of Collaborative Sensemaking and Planning Tools for Personnel Recovery". *Knowledge Media Institute Technical Report*. http://kmi.open.ac.uk/publications/index.cfm?trnumber=KMI-06-07.

Taubman, D.S. and Marcellin, M.W. (2002). *JPEG 2000: Image Compression Fundamentals, Standards and Practice*. Norwell, MA: Kluwer Academic Publishers.

Tebri, H. (2004). "Formalisation et Spécification d'un Système de Filtrage Incrémental d'Information", *PhD Thesis*. Université Paul Sabatier de Toulouse, France.

Tejada, S., Knoblock, C. and Minton, S. (2001). "Learning Object Identification Rules for Information Integration", *Information Systems*, 26(8): 607–633.

Teranishi, Y., Kamahara, J. and Shimojo, S. (2006). "MapWiki: A Ubiquitous Collaboration Environment on Shared Maps", *International Symposium on Applications and the Internet Workshops*. Phoenix, AZ, 146–149.

Tochtermann, K., Riekert, W.-F., Wiest, G., Seggelke, J. and Mohaupt-Jahr, B. (1997). "Using Semantic, Geographical, and Temporal Relationships to Enhance Search and Retrieval in Digital Catalogs", *1st European Conference on Research and Advanced Technology for Digital Libraries (LNCS, Vol. 1324)*. Pisa, Italy, 73–86.

Torres, M. (2002). "Semantics Definition to Represent Spatial Data", *International Workshop on Semantic Processing of Spatial Data* (Geopro-2002). Mexico City, Mexico.

Toyama, K., Logan, R. and Roseway, A. (2003). "Geographic Location Tags on Digital Images", *11th ACM International Conference on Multimedia*. New York: ACM Press, 156–166.

Tummarello, G., Morbidoni, C. and Nucci, M. (2006). "Enabling Semantic Web Communities with DBin: An Overview", *International Semantic Web Conference*. Athens, GA.

Tummarello, G., Morbidoni, C., Petersson, J., Puliti, P. and Piazza, F. (2004). "RDFGrowth, a P2P Annotation Exchange Algorithm for Scalable Semantic Web Applications", *International Workshop on Peer-to-Peer Knowledge Management (P2PKM-2004)*. Boston, MA.

Tummarello, G., Morbidoni, C., Puliti, P. and Piazza, F. (2005). "Signing Individual Fragments of an RDF Graph", *14th International World Wide Web Conference (WWW-2005)*. Chiba, Japan.

Turner, A. (2006). *Introduction to Neogeography.* Sebastopol, CA: O'Reilly Press.

U.S. Census Bureau (1990). *Frequently Occurring First Names and Surnames from the 1990 Census*. http://www.census.gov/genealogy/names/.

Udell, J. (2005). Annotating the Planet with Google Maps. *InfoWorld*. March 4, 2005: http://www.infoworld.com/article/05/03/04/10OPstrategic_1.html.

UNECE (United Nations Economic Commission for Europe Convention on Long Range Transboundary Air Pollution) (2005). "Monitoring of Atmospheric Heavy Metal Deposition in Europe Using Bryophytes. Experimental Protocol 2005/2006 Survey". Bangor, UK: UNECE ICP Vegetation.

United States Department of the Interior (2003). *Geospatial One-Stop Website*. http://www.geodata.gov/.

United States Geological Survey (2004). *National Earthquake Information Center Website*. http://neic.usgs.gov/.

United States National Research Council (1999). *Distributed Geolibraries: Summary of a Workshop*. http://www.nap.edu/html/geolibraries/.

USGS (2005). *Geographic Names Information System (GNIS)*. U.S. Board on Geographic Names. http://geonames.usgs.gov/domestic/index.html.

Vahidov, R. (2005). "Intermediating User-DSS Interaction with Autonomous Agents", *IEEE Transactions on Systems, Man and Cybernetics,* 35(6): 964–970.

Vandeloise, C. (1986). *L'espace en Français: Sémantique des Prépositions Spatiales*. Paris: Editions du Seuil.

Vasudevan, V. (2001). *A Web Services Primer*. http://webservices.xml.com/pub/a/ws/2001/04/04/webservices/index.html.

Vretanos, P.A. (2005). "Web Feature Service (WFS) Implementation Specification 1.1", *Open Geospatial Consortium.* http://portal.opengeospatial.org/-files/?artifact_id=8339.

W3C (2002). *Describing and Retrieving Photos Using RDF and HTTP*. World Wide Web Consortium. http://www.w3.org/TR/photo-rdf/.

W3C (2003). *Basic Geo (WGS84 lat/long) Vocabulary*. World Wide Web Consortium. http://www.w3.org/2003/01/geo/.

W3C (2004). *OWL Web Ontology Language Guide, W3C Recommendation 10 February 2004*. World Wide Web Consortium. http://www.w3.org/TR/owl-guide/.

W3C (2006a). *SPARQL Protocol and RDF Query Language*, World Wide Web Consortium. http://www.w3.org/TR/rdf-sparql-query/.

W3C (2006b). *OWL Web Ontology Language Overview*. World Wide Web Consortium. http://www.w3.org/TR/owl-features/.

W3C-Cal (2002). *RDF Calendar Workspace,* World Wide Web Consortium. http://www.w3.org/2002/12/cal/.

W3C-Exif (2003). *Exif Vocabulary Workspace – RDF Schema*, World Wide Web Consortium. http://www.w3.org/2003/12/exif/.

W3C-RDF (2002). *Resource Description Framework,* World Wide Web Consortium. http://www.w3.org/RDF/.

Wang, C., Xie, X., Wang, L., Lu, Y. and Ma, W.-Y. (2005). "Detecting Geographic Locations from Web Resources", *2nd International Workshop on Geographic Information Retrieval (GIR-2005)*. Bremen, Germany, 17–24.

Want, R. (2006). "An Introduction to RFID Technology", *IEEE Pervasive Computing*, 5(1): 25–33.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis.* Cambridge: Cambridge University Press.

Weiner, D., Harris, T.M. and Craig, W.J. (2002). "Community Participation and Geographic Information Systems", *Community Participation and Geographic Information Systems.* Eds. W. J. Craig et al. New York: Taylor and Francis, 3–16.

Weiss, S.M., Indurkhya, N., Zhang, T. and Damerau, F.J. (2005). *Text Mining – Predictive Methods for Analyzing Unstructured Information*. New York: Springer.

Weissenberg, N., Voisard, A. and Gartmann, R. (2004). "Using Ontologies in Personalized Mobile Applications", *International Symposium on Geographical Information Systems*. Washington, DC, 2–11.

Westermann, U. and Jain, R. (2006). "Events in Multimedia Electronic Chronicles (E-Chronicles)", *International Journal on Semantic Web and Information Systems*, 2(2): 1–23.

White, S. and Smyth, P. (2003). "Algorithms for Estimating Relative Importance in Networks", *Proceedings 9th Conference of the ACM Special Interest Group on Knowledge Discovery in Data Mining,* SIGKDD. Washington, DC, 266–275.

Wick, J.V., Callas, J.L., Norris, J.S., Powell, M.W. and Vona, M.A. (2005). "Distributed Operations for the Mars Exploration Rover Mission with the Science Activity Planner", *Aerospace IEEE Conference,* 4162–4173.

Widlöcher, A. and Bilhaut, F. (2005). "La Plate-forme Linguastream: Un Outil d'Exploration Linguistique sur Corpus", *12th Conférence Traitement Automatique du Langage Naturel*. Dourdan, France.

Wilk, C. (2005). "Welt in Händen: Arbeiten mit Google Earth und World Wind", *iX – Magazin für professionelle Informationstechnik,* 12/05: 50–62.

Wise, J.A. (1999). "The Ecological Approach to Text Visualization", *Journal of the American Society for Science,* 50(9): 814–835.

Wood, J. (2005). "'How Green Is My Valley?' Desktop Geographic Information Systems as a Community-based Participatory Mapping Tool", *Area*, 37(2): 159–170.

Wood, M. (2003). "Some Personal Reflections on Change, the Past and Future of Cartography", *The Cartographic Journal*, 40(2): 111–115.

Woodruff, A.G. and Plaunt, C. (1994). "GIPSY: Georeferenced Information Processing System", *Journal of the American Society for Information Science*, 45(9): 645–655.

Worboys, M. (2005). "Event-oriented Approaches to Geographic Phenomena", *International Journal of Geographical Information Science,* 19(1): 1–28.

WSML (2005). The Web Service Modeling Language WSML (D16.1v0.21). http://www.wsmo.org/TR/d16/d16.1/v0.21/.

WSMO (2005a). WSMO Primer (D3.1v0.1). http://www.wsmo.org/TR/d3/d3.1/.

WSMO (2005b). *A Conceptual Comparison between WSMO and OWL-S (D4.1v0.1)*. http://www.wsmo.org/TR/d4/d4.1/v0.1/.

Xu, C. and Cheung, S.C. (2002). "Inconsistency Detection and Resolution for Context-aware Middleware Support", *International Symposium on the Foundations of Software Engineering*. Lisbon, Portugal: ACM Press, 336–345.

Yangarber, R. and Grishman, R. (1998). "NYU: Description of the Proteus/PET System as Used for MUC-7 ST", *7th Message Understanding Conference*. Fairfax, VA: http://www-nlpir.nist.gov/related_projects/muc/.

Zipf, G.K. (1949). *Human Behaviour and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley.

Zong, W., Wu, D., Sun, A., Lim, E. and Goh, D.H. (2005). "On Assigning Place Names to Geography Related Web Pages", *5th ACM/IEEE-CS Joint Conference on Digital Libraries (*JCDL-2005). Denver, CO: ACM Press, 354–362.

Online Resources

1. NASA World Wind ▪ worldwind.arc.nasa.gov
2. Google Earth ▪ earth.google.com
3. Microsoft Virtual Earth 3D ▪ maps.live.com
4. Web Map Service (WMS) ▪ www.opengeospatial.org/standards/wms
5. Open Geospatial Consortium (OGC) ▪ www.opengeospatial.org
6. Blue Marble ▪ earthobservatory.nasa.gov/newsroom/bluemarble
7. Landsat 7 ▪ landsat.gsfc.nasa.gov
8. MetaCarta ▪ www.metacarta.com
9. GeoTagging Flickr Group ▪ www.flickr.com/groups/geotagging
10. Moderate Resolution Imaging Spectroradiometer ▪ modis.gsfc.nasa.gov
11. SigAlert Mash-Up ▪ bbs.keyhole.com/ubb/download.php?Number=75329
12. WHOIS Lookup ▪ www.whois.net
13. Geographic Names Information System ▪ geonames.usgs.gov
14. World Gazetteer ▪ www.world-gazetteer.com
15. UN Statistics Division: Geographical Names ▪ un-stats.un.org/unsd/geoinfo
16. Getty Thesaurus ▪ www. getty.edu/research/conducting_research/vocabularies/tgn
17. ISO 3166 Maintenance Agency ▪ www.iso.org/iso/en/prods-services/iso3166ma/index.html
18. Dublin Core Metadata Initiative ▪ www.dublincore.org
19. MicroPatent Aureka Software Suite ▪ www.cartia.com
20. IDIOM Research Project ▪ www.idiom.at
21. Associated Press Mash-Up ▪ www.81nassau.com/apnews
22. Associated Press RSS News Feeds ▪ hosted.ap.org/dynamic/fronts/RSS
23. Yahoo! Geocoding API ▪ developer.yahoo.com/maps/rest/V1/geocode.html
24. KeyHole ▪ www.keyhole.com
25. GeoTango ▪ www.geotango.com
26. Vexcel ▪ www.vexcel.com
27. MS Flight Simulator ▪ www.microsoft.com/games/flightsimulator
28. Geography Markup Language (GML) ▪ www.opengeospatial.org/standards/gml
29. University of Minnesota Map Server ▪ mapserver.gis.umn.edu
30. GeoTIFF Format Specification ▪ www.remotesensing.org/geotiff
31. Flickr ▪ www.flickr.com
32. Plazes ▪ beta.plazes.com
33. Upcoming.org ▪ upcoming.org
34. Microformats ▪ microformats.org
35. Google Keyhole Markup Language ▪ earth.google.com/kml
36. Placeopedia ▪ www.placeopedia.com
37. Mappr ▪ www.mappr.com
38. Open Directory Project ▪ dmoz.org
39. Google Maps ▪ maps.google.com

[40] Yahoo! Maps ▪ maps.yahoo.com
[41] MSN Live ▪ local.live.com
[42] Geocoder.us ▪ geocoder.us
[43] Google SOAP Search API ▪ code.google.com/apis/soapsearch
[44] Tiger Data Set from U.S. Census Bureau ▪ tiger.census.gov
[45] LSDIS Spatiotemporal Data Sets ▪ lsdis.cs.uga.edu/projects/semdis/spatiotemporal
[46] RDF Gravity ▪ semweb.salzburgresearch.at/apps/rdf-gravity
[47] NekoHtml Parser ▪ java-source.net/open-source/html-parsers/nekohtml
[48] Geographic Services from Yahoo ▪ developer.yahoo.com/maps
[49] JUNG Java Library for Graph Visualization ▪ jung.sourceforge.net
[50] Finding Semantic Associations Online ▪ lsdis.cs.uga.edu:8080/SemanticAssociationDemo
[51] SemDis API ▪ lsdis.cs.uga.edu/projects/semdis/api
[52] Spatiotemporal REST Services ▪ lsdis.cs.uga.edu:8080/SemDisServices
[53] Geo RDF Initiative ▪ www.w3.org/2003/01/geo
[54] Cordial ▪ www.synapse.com
[55] Institut Géographique National (IGN) ▪ www.ign.fr
[56] Via Michelin ▪ www.viamichelin.fr
[57] PostGIS ▪ postgis.refractions.net
[58] eXist ▪ exist.sourceforge.net
[59] Cross-Language Evaluation Forum (CLEF) ▪ www.clef-campaign.org
[60] GeoRSS 1.0 ▪ www.georss.org/1
[61] Web Feature Service (WFS) ▪ www.opengeospatial.org/standards/wfs
[62] DartMaps ▪ dartmaps.mackers.com
[63] ChicagoCrime ▪ www.chicagocrime.org
[64] OpenStreetMap: The Free Wiki World Map ▪ www.openstreetmap.org
[65] Platial ▪ www.platial.com/splash
[66] Tracks4Africa ▪ www.tracks4africa.co.za
[67] Amazon Conservation Team ▪ www.amazonteam.org
[68] Zooomr ▪ zooomr.com
[69] Picasa ▪ picasaweb.google.com
[70] Flickr Map ▪ www.flickr.com/map
[71] Zoto ▪ www.zoto.com
[72] Jpgearth ▪ www.jpgearth.com
[73] Loc.alize.us ▪ loc.alize.us
[74] Greasemonkey ▪ greasemonkey.mozdev.org
[75] Sharing Places ▪ www.sharing-places.com
[76] Suunto ▪ www.suunto.com
[77] Joseki ▪ www.joseki.org
[78] Jena ▪ jena.sourceforge.net
[79] Jakarta Velocity ▪ jakarta.apache.org/velocity
[80] GeoOntologies ▪ www.mindswap.org/2004/geo/geoOntologies.shtml

[81] Spatial Ontologies ▪ space.frot.org/ontology.html
[82] Geographic Ontologies ▪ loki.cae.drexel.edu/~wbs/ontology/iso-19115.htm
[83] GeoOnion ▪ esw.w3.org/topic/GeoOnion
[84] Photography Vocabulary ▪ www.wasab.dk/morten/2003/11/photo
[85] Camera OWL Ontology ▪ www.xfront.com/camera/camera.owl
[86] SKOS Core Vocabulary Specification ▪ www.w3.org/TR/swbp-skos-core-spec
[87] RDF Semantics ▪ W3C Recommendation ▪ www.w3.org/TR/rdf-mt
[88] OWL Web Ontology Language Overview ▪ www.w3.org/TR/owl-features
[89] Welkin ▪ simile.mit.edu/welkin
[90] Longwell ▪ simile.mit.edu/longwell
[91] AIR World Wide ▪ www.air-worldwide.com
[92] EQECAT ▪ www.eqecat.com
[93] Risk Management Solutions (RMS) ▪ www.rms.com
[94] Autrian Portal ▪ geoinfo.lfrz.at/website/egisroot/services/ehora2/viewer.htm
[95] Geoportail ▪ www.geoportail.fr
[96] Geocatalogue ▪ www.geocatalogue.fr
[97] United States Geological Survey (USGS) ▪ seamless.usgs.gov
[98] PRIM.NET ▪ www.prim.net
[99] Kheops Technology ▪ www.kheops-tech.com
[100] AirBase ▪ airbase.eionet.europa.eu
[101] EEA Ozone Web ▪ www.eea.europa.eu/maps/ozone
[102] Homeland Security Working Group ▪ www.fgdc.gov/HSWG
[103] Sensor Web Alliance ▪ www.sensorweb-alliance.org
[104] Web Service Modelling eXecution environment (WSMX) ▪ www.wsmx.org
[105] SWING ▪ www.swing-project.org
[106] Web Processing Service (WPS) ▪ www.opengeospatial.org/standards/requests/28
[107] OWL-S 1.0 Release ▪ www.daml.org/services/owl-s/1.0
[108] Amstel Botel Amsterdam ▪ www.amstelbotel.nl
[109] Google Maps Mania ▪ googlemapsmania.blogspot.com
[110] Universal Description, Discovery, and Integration (UDDI) ▪ www.uddi.org
[111] Web Services Description Language (WSDL) ▪ www.w3.org/TR/wsdl
[112] Simple Object Access Protocol (SOAP) ▪ www.w3.org/TR/soap12-part1
[113] Jabber Protocol ▪ www.jabber.org
[114] Google Web Toolkit ▪ code.google.com/webtoolkit
[115] eMerges ▪ irs-test.open.ac.uk/sgis-dev
[116] Firefox Browser ▪ www.mozilla.com/firefox
[117] Mapufacture GeoRSS Aggregator ▪ mapufacture.com
[118] GeoNames ▪ www.geonames.org
[119] ArcGIS Explorer ▪ www.esri.com/software/arcgis/explorer
[120] AKTive Response ▪ www.e-response.org/AKTiveResponse
[121] SFSW2006 ▪ www.semanticscripting.org/SFSW2006
[122] Semantic Web Challenge ▪ challenge.semanticweb.org

Index

adjacency 7, 96-99, 118
aerial imagery 11, 155
agriculture 181, 211, 227
air
  pollution 191
  quality 35, 201-208
airbase 257, 289
ambiguity 6f., 20, 61, 82, 117-122, 127
  geo/geo 61, 288
  geo/non-geo 61, 108f., 119
analysis
  comparative 149
  content 3
  structural 152
  textual 115
anchor theory 41, 44
annotation 5-7, 11, 43, 131, 137, 162-164, 169-176, 229f., 233-235, 284
  automated 6
  geospatial 11, 175
  manual 55
application development 225-227
Application Programming Interface (API) 5, 53, 62, 71-74, 87, 248, 254, 282, 288
ArcGIS 256, 274, 289
assessment ix, x, xiv, 61f., 179-192, 199-202, 208, 237, 264f., 269-273, 278-281
Asynchronous JavaScript and XML (AJAX) 112, 226, 254
avatar 4
axiom 238
bandwidth 27f., 36, 224f.
Bayes 123, 126f.
bioaccumulation 191-194, 278
bitplane 30f., 37
blog (Web log) 9, 105-107, 110f., 114, 270
Blue Marble 4, 287, 291
bookmarking 111, 114
business
  logic 224
  model 3
cartographic
  data vii, 155
  visualization 104, 209-214
centrality 142f., 146-149, 259, 267, 280
choreography 249
citation 141, 279
classification 7, 23, 56, 83, 114, 117, 122f., 143, 169, 227, 278
clustering 8, 59, 145f., 161f., 193, 270, 277, 280
clutter 10
cognitive mapping 209
co-kriging 202-205
collaborative
  environment 78
  mapping 153, 155-158
collective intelligence 14
community
  empowerment 153f., 156-158
  geospatial 27, 227
  local 154, 181
  mapping 153, 157f.
  member 154, 157
compression 27-31, 36f., 274, 283
conceptual
  fuzziness 44
  search 47-54
  space 237, 252
content
  analysis 3
  management 3, 11, 184
  model 18, 224, 228
  production vii, 9f.
  syndication 107
  user-generated 11, 157
context
  information 5, 67, 72-78, 131-136, 252
  map 131, 134-139
co-occurrence 119, 241-243
copyright 156, 187, 228
corpus 93-98, 101-104, 108f., 119, 127, 261, 273, 285
crawler 12, 58f., 62-65
cross-validation 204
cultural
  context 6
  factor 228
  heritage 93, 97f., 104, 171f.
data
  cartographic vii, 155
  fusion 217, 220
  management 14, 38, 225f.
  meteorological 202, 208
  model 67-70, 73, 88, 124f., 182, 186, 218
  quality 226
  raster 28, 35-38, 180
database
  indicator 188
  management 194f.
  relational 19, 40, 124, 184
  spatial 18, 28, 37, 40f.
  synchronization 20
  system 37, 59, 196

decision
making 16, 26, 132, 139, 153-155, 158, 184, 209, 212f., 229-234
support 171, 176, 209, 215, 249
tree 123, 126
decoder 30f., 36
description logic 40, 237f.
digital
atlas 185f., 190
elevation model 107, 180
library 95, 109, 117
photography 136, 159f.
sampling protocol 197, 200
disambiguation ix, 5, 7, 43, 60f., 79-83, 88f., 108-110, 117-128
distributed
collaboration 139
environment 159, 163f., 167, 229
dmoz 56, 60, 65, 114, 287
document
collection 93, 104, 110, 117f., 125-127
HTML 194-196
structured 117
domain ontology 9, 218f., 232f., 250
Dublin Core 7, 56, 164, 168f., 238, 287, 291
earthquake 111, 284
ecoregion 191, 193f., 200, 278
ecosystem 35, 191f., 200f., 215
emergency 15, 17, 172, 183, 209, 211, 214, 247-256, 264, 272-274
emission level vii
encyclopedia vii, 8f., 272
environment
collaborative 78
environmental
data 77
indicator v-vii
information 39
management 24, 211
monitoring 192, 196-200
protection 16
epidemiology 141
exposure assessment 179f., 203
Extensible Markup Language (XML) 5, 17-21, 33, 37, 48f., 56, 70, 77, 94-107, 113, 168, 173-175, 187, 224-226, 239, 255-257, 261
Feature Portrayal Service (FPS) 20f., 25
feature space 117, 123
fire department 17, 20, 24
first-mover advantage 11
Flickr 55, 160-163, 169, 287f., 291
focused
crawler 55, 58-65
Web search 56
force-directed placement 8
forestry 24, 223
gazetteer 7, 56, 60, 95, 100, 104-106, 109, 117-125, 134, 269f., 287, 291
generalization 212f., 245, 252
Geo RDF 288
geocoding 6, 7, 55-58, 64, 79, 82, 87, 105-107, 112, 159, 181, 184, 288
Geodata Infrastructure (GDI) 223, 227f.
geographic
core model 94-97, 102-104
data v, 13-15, 24-27, 57, 79f., 83, 114f., 119, 179-196, 200, 213, 223-233, 256, 284
entity 57, 60f., 80, 87, 117f., 124-126
information extraction 93, 180
information retrieval 57, 93, 117
information system (GIS) xiii, 10, 22-28, 31, 40-42, 95-97, 101f., 117, 132-142, 153-157, 176-200, 209-216, 223-228, 250f., 257-286
knowledge base 117-127
named entity 118, 122-127
relationship 117-127
scope 6
stop word 122
Geography Markup Language (GML) 7, 15-28, 32-38, 48, 56, 89, 155, 185, 218, 263, 267, 277, 287, 291
geological survey 136
geometric shape 100, 137
geometry 18, 24, 41, 47f., 80f., 117, 121, 163, 233, 263, 268
geoparsing 6, 57-65
geo-referenced
data vii, 56, 162
image 27, 33
news 10
GeoRSS 106-109, 112, 288f.
geospatial
annotation 11, 175
context vii, 3-7, 11, 55, 115, 139
literacy vii, 26
platform v-vii, 3, 11-13, 106, 182
search 58-60
geotagging vii, ix, 3-6, 11-13, 55f., 110, 159-163, 170-172, 257, 287, 291
GeoTIFF 29, 32-36, 287, 291
Getty Thesaurus of Geographic Names (TGN) 7, 117, 268
GIScience 142-147, 150-154, 268-272
Global Positioning System (GPS) 4f., 68-77, 115, 133-138, 157-163, 166f., 181-188, 227
Google Earth vii, 3-5, 11-13, 26, 39, 56, 67-78, 106, 109, 139, 153-163, 175, 227, 285-291
Google Maps 10, 13, 39, 56, 68, 73-79, 87, 106, 111, 153-163, 167, 175, 254, 265, 284-289
Google Web Toolkit 254, 289
government agency v, 24f., 115
graph
analysis 57
layout 141
model 174

hazard
assessment 179
data 181, 187-189
exposure 179-184, 190
natural 179-185, 190
heterogeneity 143, 148-151, 200, 213, 218, 229
homeland security 25
homonym 43, 44
hurricane 3-5, 106, 111, 153, 183, 266
hydrology 43, 181
Hypertext Markup Language (HTML) 56, 68, 117, 136, 184, 194-196, 263
impact assessment 201, 207
index file 100f.
indexer 55, 58
indicator database 188
information
extraction 93f., 97, 102-109, 114f.
retrieval 4, 94-98, 122, 127, 224-237, 253
service vii, 4, 11, 39, 78, 207f.
system 94, 97, 115, 132-134, 189-194, 201, 227
insurance ix, 39, 179-190, 261
interactivity viii, 3, 9f., 36, 111, 158, 184, 196f., 198, 208-210, 226, 260, 263, 270, 280
interface
metaphor 4
query 8, 65
technology v
user v, 4, 14, 74, 88, 104, 110, 162, 170, 176, 196
International Standardization Organization (ISO) 7, 17, 29, 33-36, 182, 187-190, 213, 225, 233, 270, 277, 287, 291
interoperability xii, 17, 20, 27f., 35, 155, 174f., 183, 187, 190, 194, 200f., 216, 223-232, 238, 247-249, 261, 273, 280f.
interpolation x, 201-208, 258, 264-266, 270
intersection 39-42, 51, 93-95, 101, 104, 144-146, 187, 219, 227f., 238, 242, 252
invocation 48, 53, 71, 88, 229f., 233f., 248-250
Jena Semantic Web Framework 168, 288
JPEG 2000 ix, 27-38, 270, 277, 283
Keyhole Markup Language (KML) 11, 75, 287
knowledge
acquisition 199
planet 8f.
repository v-vii, 3, 9f., 13f., 215
representation 237
kriging 202-205
Landsat 7 226, 287, 291
legacy system 224, 248
linear regression 202, 205f.
linguistic
analysis 120
processing 98f.
Loc.alize.us 161-163, 288
localized search 66
location-based
application 55
indexing 58
information retrieval 55
search 55, 63-65
service (LBS) 6, 11, 14, 184
logistic regression 123
loosely coupled 115, 216, 220, 224
machine learning 117-127
management
content 3, 11, 184
environmental 24, 211
map
context 131, 134-139
topographic vii, 133-135, 181
mapping
cognitive 209
collaborative 153-158
Mappr 56, 287
MapQuest 13, 156
mash-up 5, 11, 73, 109, 155f., 176, 226, 247, 256, 270
maximum likelihood estimation 123
media coverage vii, 12-14
mediator 250, 255
metadata vii, 6f., 10, 13, 23-35, 56, 79f., 86, 89, 110-113, 131-136, 159-199, 209, 216, 229-235, 262-264, 276-281, 287, 291
Microsoft Live Local vii, 39
Microsoft Virtual Earth 11, 26, 106, 156f., 287, 291
middleware 68f., 73, 263, 286
minimum bounding rectangle 100
mobile
device 55, 74, 153, 173
phone 12
modeling
approach 141, 144, 152
framework 151
language x, 217, 230-238, 263, 286
object-oriented 124
Moderate Resolution Imaging Spectroradiometer (MODIS) 5, 112, 216f., 287, 291
monitoring
data 191-196, 201-205
network 198-200
site 192, 198-201
municipality 184-188
museum 55, 60, 77, 93
MySQL 184, 196, 273
named entity 6, 61, 94f., 103, 117-122
NASA World Wind v-vii, 3, 5, 139, 287, 291
natural
event 5, 181
hazard 179-185, 190
language processing 6, 98, 105f., 109, 114-120, 127
resource 211, 227
network
analysis 142
connection 28, 37, 74
effect 10f., 14, 163
infrastructure 57

road 17, 34, 155, 252
society vii
news
aggregator vii, 9, 107
coverage 3, 10-12
media vii, 9
service 9
newsroom 9f., 281, 287, 291
normalization 52, 61, 146, 240f., 273
ontology xii, 7, 40-54, 72, 78, 86, 95, 110, 168-176, 215-220, 230-259, 263, 267-289
Open Directory Project (dmoz) 56, 60, 65, 114, 287
Open Geospatial Consortium (OGC) 7, 17, 22-28, 32-38, 48, 80, 111, 155, 182-196, 213-234, 259, 264, 277-279, 284, 287, 291
open-source v, 5, 9, 155, 176, 191-196, 234, 288
orchestration 249-251
ozone vii, 201-208, 259f., 265, 289
part of speech 99, 106-109, 260
participatory information system ix, 153f.
pattern matching 12, 72-78, 109f.
pervasive computing 67, 70-72, 78
photo
collection v, 5, 159, 164-170
sharing 56, 159f., 163
Picasa 160-163, 288
planetary exploration 131
police 16f., 20, 253
political campaign 14
population 42, 106f., 111, 119, 181, 185, 188, 203-208, 226, 231, 271
PostGIS 70-75, 195f., 288
PostgreSQL 74f., 195-197
precipitation 203f.
precision 29, 57, 63, 66, 100-103, 127, 133, 226, 231, 243-245, 249
proceedings 9, 145, 262, 271, 276-278, 285
programming language 109, 225
proximity 83-89, 96, 157, 163, 215, 253, 267
query
definition 52
interface 8, 65
processing 58, 79
result 41, 48, 51-54, 104
term 7, 10
Radio Frequency Identification (RFID) 4f., 67-78, 259, 274, 285
rainforest 157
raster data 28, 35-38, 180
Really Simple Syndication (RSS) ix, 77, 105-114, 159, 287, 291
reasoning
qualitative 40, 44
quantitative 40
recall 63, 102f., 127, 231
regression
linear 202, 205f.
logistic 123
regular expression 12, 72
relevance ranking 7, 110
remote sensing 133
research community 141, 145-152, 234
Resource Description Framework (RDF) 54-56, 79-89, 163-176, 260f., 270, 276-289
road network 17, 34, 155, 252
satellite
image 12, 27, 105-107, 115, 157, 180, 251
observation 180, 183
scientific
collaboration 131f.
data 131-135, 139
journal 8
survey 131f.
search
assisted 61
conceptual 47-54
localized 66
search engine 7, 11f., 39, 55-65, 79, 85, 106-110, 115, 150, 187
semantic
annotation 6, 172, 219, 229f., 233f.
association 85-87
dimension 79, 141
discovery 229, 234
enrichment v, 216, 218
network 141-147
processing 103
Web ix-xiv, 39-54, 72, 79, 88f., 114, 141, 163-176, 219, 229-238, 247f., 256-272, 279-289
semantics vii, 39, 41, 44, 66, 79-104, 141, 217, 228-239, 247-251, 258, 262, 269-289
service broker 235
service-oriented architecture (SOA) v, 47f., 52-54, 183, 216-218, 224, 227, 233-235, 245
servlet 75, 188, 254
significance 57, 101f., 213
similarity
estimation 244f.
function 240
measurement 142f., 150-152, 235-242
Simple Object Access Protocol (SOAP) 233-235, 248, 288f.
social
bookmarking 114
change 13
network v, 3, 141, 145, 151f., 156-158, 163, 169, 247
software agent 70, 76, 105-107, 132, 139, 215, 220
source code v, 48, 194f.
SPARQL Query Language 48, 166-168, 279, 285
spatial
context 55, 211, 216, 247, 253
data infrastructure (SDI) 15, 25, 187, 190, 209-211, 259, 262, 268, 272, 277
dimension 84f., 208
indexing 55, 87, 110
interpolation 202
network 141f.

query 41, 75, 194
relationship 39-44, 66, 95, 110, 152, 164-179
search 55-59, 63, 66
trend 200
statistical
analysis 119f., 199
indicator 185, 202
model 119f., 123
stop word 103, 122
subsumption 235f., 240, 245
supply chain 15, 219
syntax 47, 118, 168, 228, 232, 238f., 258, 281
tagger 94, 99, 108f., 122, 260
target audience 3, 10f.
taxonomy 7, 23, 235
temporal
dimension 9, 79, 181
proximity 85
trend 191-193
text analysis 115
tightly coupled 224
topic category 187
topographic map vii, 133-135, 181
topology 8, 18, 40, 50, 79-81, 86, 95f., 100, 117, 233, 242, 262-265, 274-276, 280f.
toponym ix, 96, 106-108, 117-128, 267, 273, 282
traffic
accident 16f., 41
management 17, 184
statistics 41
transcoder 30
tuple space 73-77
uncertainty 41-44, 189, 208
United States Geological Survey (USGS) 29, 107, 111, 137f., 183f., 289
Universal Description, Discovery and Integration (UDDI) 233-235, 248, 289
user
authorization 200
interface v, 4, 14, 74, 88, 104, 110, 162, 170, 176, 196
profile v, 54, 219
user-generated content 11, 155-157
vagueness 41f., 44
vector space 123
vegetation 111, 191-193, 200, 219, 260, 284
virtual
globe 157
machine 132
visual perception 214
visualization ix, 19f., 56, 73, 87, 97, 102-106, 115, 139f., 152, 172, 176, 182, 186, 194, 209-230, 260f., 267-272, 280, 285, 288
waterway 236, 239, 244
wavelet 28, 30f., 37
weather
data 180
report 24
Web
community 17, 67f., 159
coverage service (WCS) 28, 35-37, 218
feature service (WFS) 4, 19-25, 35f., 48-53, 155, 184-190, 218, 227, 231-233, 284, 288
information retrieval 55, 57
map service (WMS) 4, 15, 20f., 35f., 111f., 155, 184-196, 218, 225-233, 287, 291
ontology language (OWL) 7, 40-48, 51-56, 172-174, 218, 235-237, 270, 275, 282-289
page 6, 40, 53-66, 98, 105-117, 141-145, 253
processing service (WPS) 231, 233, 289
registry service (WRS) 22-25
resource 5-7, 11, 55f., 114
search 55f., 64, 79, 110, 155
server 4, 57, 75, 155, 188, 194f.
service v, 15, 18f., 23, 27, 35, 41, 48-54, 68, 71, 79, 87, 93, 97, 104, 107, 132f., 137, 182, 215-217, 227-238, 249-255
service modeling ontology (WSMO) 218, 230-238, 245-253, 266, 279, 286
services description language (WSDL) 233-235, 248, 289
site vii, 6, 12, 65f., 70, 80f., 87, 105-114, 141-146, 156-160, 184, 227, 233, 245
Web 2.0 14, 55, 68, 114, 155-158, 223-228, 247, 251, 278
Web log (blog) 9, 105-107, 110f., 114, 270
WebGIS ix, 191f., 194-200, 272
Whois Lookup 287, 291
Wiki vii, 9, 107, 288
wildfire 111-115, 183
workflow 3, 10, 17, 27-38, 139, 218, 248
World Wide Web Consortium (W3C) 7, 17, 40, 56, 163-172, 225, 269f., 275-289
XML Schema 18, 239
XSL Transformations (XSLT) 48
Yahoo!
Geocoding API 10, 81, 87, 287, 291
Local 39
Maps 56, 156, 160, 163, 288