LNCS 6631

Christophe Claramunt
Sergei Levashkin
Michela Bertolotto (Eds.)

GeoSpatial Semantics

4th International Conference, GeoS 2011
Brest, France, May 2011
Proceedings

Springer

Lecture Notes in Computer Science 6631

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Christophe Claramunt Sergei Levashkin
Michela Bertolotto (Eds.)

GeoSpatial Semantics

4th International Conference, GeoS 2011
Brest, France, May 12-13, 2011
Proceedings

Springer

Volume Editors

Christophe Claramunt
Naval Academy Research Institute
29240 Brest Cedex 9, France
E-mail: christophe.claramunt@ecole-navale.fr

Sergei Levashkin
Instituto Politecnico Nacional
Centro de Investigacion en Computacion
07738 Mexico City, Mexico
E-mail: sergei@cic.ipn.mx

Michela Bertolotto
University College Dublin
School of Computer Science and Informatics
Dublin, Ireland
E-mail: michela.bertolotto@ucd.ie

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-20629-0 e-ISBN 978-3-642-20630-6
DOI 10.1007/978-3-642-20630-6
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011925395

CR Subject Classification (1998): J.2, J.4, H.2.8, H.3, H.2, H.4, I.2.9

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The fourth edition of the International Conference on Geospatial Semantics (GeoS 2011) was held in Brest, France, during May 12–13, 2011.

Geospatial semantics (GEOS) is an emerging research area in the domain of geographic information science. It aims at exploring strategies, computational methods, and tools to support semantic interoperability, geographic information retrieval, and usability. Research on geospatial semantics is intrinsically multidisciplinary and therefore GeoS traditionally brings together researchers whose expertise will address issues from diverse fields such as cognitive science, geography, linguistics, mathematics, philosophy, and information technology.

The fourth edition of GeoS provided a forum for the exchange of state-of-the-art research results in the areas of modelling and processing of geospatial semantics. Research in geospatial semantics is critical for the development of next-generation spatial databases and geographic information systems, as well as specialized geospatial Web services. Within the context of the Semantic Web, the need for semantic enablement of geospatial services is crucial, given the ever-increasing availability of mainly unstructured geospatial data. This problem is exacerbated by the very recent and ever-growing phenomena of crowd sourcing and volunteered geographic information.

These proceedings contain full research papers which were selected from among 23 submissions received in response to the Call for Papers. Each submission was reviewed by three or four Program Committee members and 13 papers were chosen for presentation. The papers focused on formal and semantic approaches, time and activity-based patterns, ontologies, as well as quality, conflicts, and semantic integration. Overall, a wide range of research efforts were presented by researchers from institutions in Italy, Mexico, France, Belgium, Japan, UK, USA, Switzerland, Portugal, and Germany.

We are thankful to all the people that contributed to the success of this event. The members of the Program Committee offered their help with reviewing submissions. Our thanks also go to the researchers and doctoral students of the Naval Academy Research Institute, who formed the Local Organizing Committee, and to Brest Metropole Oceane and Tecnopole Brest Iroise, who were the hosting institution and co-sponsored GeoS 2011. Other co-sponsors include the Intelligent Processing of Geospatial Information Laboratory (PIIG) of the Centre for Computing Research (CIC) of the National Polytechnical Institute (IPN) and the National Council for Science and Technology (Mexico) to which we are also grateful. This year's conference was held under the umbrella of the Safer Seas III - 2011 conference, therefore providing great opportunities for additional stimulating exchanges. Finally, we would like to sincerely thank all the authors who submitted papers to GeoS 2011, as well as our keynote speaker Max Craglia for his talk on "Interdisciplinary Interoperability for Global Sustainability Research."

May 2011

Christophe Claramunt
Sergei Levashkin
Michela Bertolotto

Organization

General Chairs

Sergei Levashkin	Centro de Investigación en Computación, Mexico City, Mexico
Christophe Claramunt	Naval Academy Research Institute, Brest, France

Program Chairs

Christophe Claramunt	Naval Academy Research Institute, Brest, France
Sergei Levashkin	Centro de Investigación en Computación, Mexico City, Mexico
Michela Bertolotto	School of Computer Science and Informatics, University College Dublin, Ireland

Program Committee

Neeharika Adabala	Microsoft Research, India
Ola Ahlqvist	Ohio State University, USA
Naveen Ashish	UC-Irvine, Irvine CA, USA
Ioan Marius Bilasco	Laboratoire d'Informatique Fondamentale de Lille, France
Roland Billen	University of Liege, Belgium
Tom Bittner	University at Buffalo, USA
Stefano Borgo	Laboratory for Applied Ontology, Italy
Boyan Brodaric	Geological Survey of Canada, Canada
Jean Brodeur	Natural Resources Council, Canada
Gilberto Camara	INPE, Brazil
Elena Camossi	JRC ISPRA, Italy
Tao Cheng	University College London, UK
Isabel F. Cruz	University of Illinois at Chicago, USA
Clodoveu Davis Jr.	Universidade Federal de Minas Gerais, Brazil
Christian Freksa	University of Bremen, Germany
Max Egenhofer	University of Maine, USA
Mauro Gaio	University of Pau, France
Anthony Galton	University of Exeter, UK
Bin Jiang	University of Gavle, Sweden
Marinos Kavouras	National Technical University of Athens, Greece

Alexander Klippel	The Pennsylvania State University, USA
Margarita Kokla	National Technical University of Athens, Greece
Dave Kolas	BBN Technologies, USA
Robert Laurini	INSA Lyon, France
Miguel R. Luaces	University of La Coruna, Spain
Sergio di Martino	University of Naples Federico II, Italy
Miguel Felix Mata Rivera	Instituto Politecnico Nacional, Mexico
Gavin McArdle	National University of Ireland, Maynooth, Ireland
Vasily Popovich	Saint Petersburg Institute for Informatics and Automation, Russia
Cyril Ray	Naval Academy Research Institute, France
Andrea Rodriguez	Universidad de Concepcion, Chile
Angela Schwering	University of Münster, Germany
Shashi Shekhar	University of Minnesota, USA
Emmanuel Stefanakis	Harokopio University of Athens, Greece
John Stell	University of Leeds, UK
Kathleen Stewart Hornsby	University of Iowa, USA
Christelle Vangenot	University of Geneva, Switzerland
Nancy Wiegand	University of Wisconsin-Madison, USA
Stephan Winter	University of Melbourne, Australia
Esteban Zimányi	Université Libre de Bruxelles, Belgium

Local Organizing Committee

Imad Afyouni	Naval Academy Research Institute, Brest, France
Marie Coz	Naval Academy Research Institute, Brest, France
Geraldine Del Mondo	Naval Academy Research Institute, Brest, France
Laurent Etienne	Naval Academy Research Institute, Brest, France
Remy Thibaud	Naval Academy Research Institute, Brest, France
Fabienne Vallée	Technopole Brest Iroise, Brest, France
Eric Vanderbroucke	Technopole Brest Iroise, Brest, France

Conference Organizing Committee

PIIG-Lab, Centro de Investigación en Computación, Mexico City, Mexico:

Miguel Martinez
Felix Mata
Nahun Montoya
Iyeliz Reyes
Gerardo Sarabia
Linaloe Sarmiento
Roberto Zagal

Table of Contents

Invited Talk

Ontologies and Gazetteers

Activity-Based and Temporal Issues

Models, Quality and Semantic Similarities

Retrieval and Discovery Methods

Inter-disciplinary Interoperability for Global Sustainability Research

Massimo Craglia[1], Stefano Nativi[2], Mattia Santoro[2],
Lorenzino Vaccari[1], and Cristiano Fugazza[1]

[1] European Commission, Joint Research Center,
Institute for Environment and Sustainability Ispra, Italy
{massimo.craglia,lorenzino.vaccari,
cristiano.fugazza}@jrc.ec.europa.eu
[2] Institute of Methodologies for Environmental Analysis of the National Research Council
(IMAA-CNR), Italy
{nativi,santoro}@imaa.cnr.it

Abstract. The implementation of the INSPIRE Directive in Europe and similar efforts around the globe to develop spatial data infrastructures and global systems of systems have been focusing largely on the adoption of agreed technologies, standards, and specifications to meet the (systems) interoperability challenge. Addressing the key scientific challenges of humanity in the 21st century requires however a much increased inter-disciplinary effort, which in turn makes more complex demands on the type of systems and arrangements needed to support it. This paper analyses the challenges for inter-disciplinary interoperability using the experience of the EuroGEOSS research project. It argues that inter-disciplinarity requires mutual understanding of requirements, methods, theoretical underpinning and tacit knowledge, and this in turn demands for a flexible approach to interoperability based on mediation, brokering and semantics-aware, cross-thematic functionalities. The paper demonstrates the implications of adopting this approach and charts the trajectory for the evolution of current spatial data infrastructures.

Keywords: Knowledge management, Semantic interoperability, Spatial Data Infrastructures.

1 Introduction

One of the most fundamental challenges facing humanity at the beginning of the 21st century is to respond effectively to the global changes that are increasing pressure on the environment and on human society. This priority is articulated by the International Council for Science (ICSU) as follows:

"Over the next decade the global scientific community must take on the challenge of delivering to society the knowledge and information necessary to assess the risks humanity is facing from global change and to understand how society can effectively

C. Claramunt, S. Levashkin, and M. Bertolotto (Eds.): GeoS 2011, LNCS 6631, pp. 1–15, 2011.

mitigate dangerous changes and cope with the change that we cannot manage. We refer to this field as 'global sustainability research'[1] *"*.

ICSU identified five scientific priorities, or Grand Challenges, in global sustainability research through a broad consultation involving over 1000 scientists from 85 countries in 2009-2010. These Grand Challenges include:

1. Developing the **observation** systems needed to manage global and regional environmental change.
2. Improving the usefulness of **forecasts** of future environmental conditions and their consequences for people.
3. Recognizing key **thresholds** or non-linear changes to improve our ability to anticipate, recognize, avoid and adapt to abrupt global environmental change.
4. Determine what institutional, economic and behavioural **responses** can enable effective steps toward global sustainability.
5. Encouraging **innovation** (coupled with sound mechanisms for evaluation) in developing technological, policy, and social responses to achieve global sustainability.

The increasing importance of linking the scientific effort necessary to underpin the sustainability agenda with innovation and sustainable economic growth is also at the heart of the European Union's Europe 2020 strategy[2], focusing on smart, sustainable, and inclusive growth.

The Global Earth Observation System of Systems (GEOSS[3]), envisioned by the group of eight most industrialized countries (G-8) in 2003 and currently half way in its 10-year implementation plan, provides the indispensable framework to integrate the earth observation efforts of the 84 GEO-members and 58 participating organisations. A major role of GEOSS is to promote scientific connections and interactions between the observation systems that constitute the system of systems, and address some of the scientific challenges identified by ICSU with a particular focus on nine societal benefit areas[4]. Such interactions also promote the introduction of innovative scientific techniques and technologies in the component observing systems. In this respect therefore the development of GEOSS can make a strategic contribution in delivering the objectives of the Europe 2020 strategy.

For these reasons the European Commission plays a very active role in developing GEOSS. This includes participating and co-chairing GEOSS Committees and the Data Sharing Task Force, and implementing important initiatives to collect and share environmental information for the benefit of the global society: the Infrastructure for Spatial Information in Europe (INSPIRE Directive), the Global Monitoring for Environment and Security (GMES) initiative, and the Shared Environmental

[1] `http://www.icsu-visioning.org/wp-content/uploads/Grand_Challenges_Nov2010.pdf`

[2] COM(2010)2020.

[3] `http://www.earthobservations.org/geoss.shtml`

[4] Disasters, Health, Energy, Climate, Agriculture, Ecosystems, Biodiversity, Water, and Weather.

Information System (SEIS). The European Commission also contributes to the implementation of the GEOSS Work Programme through research projects like EuroGEOSS[5], which are funded from its Framework Programme for Research & Development.

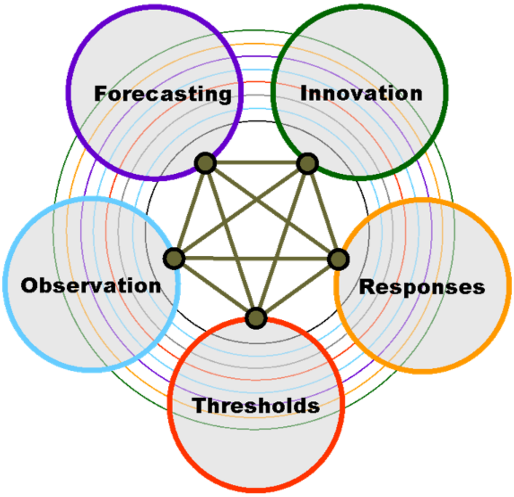


Fig. 1. The five ICSU Grand Challenges in Global Sustainability Research

The rest of the paper is organized as follows. Section 2 describes the progress and main results of the first phase of the EuroGEOSS project. Section 3 illustrates how the EuroGEOSS inter-disciplinary Initial Operating Capability (IOC) was implemented trough the adoption of a brokering approach. Section 4 presents EuroGEOSS results for advanced knowledge organization and augmented semantic search for SDIs. Finally, Section 4 concludes the paper by summarizing existing challenges, the EuroGEOSS solution proposed to address these challenges and the next steps of the project.

2 Progress and Main Results to Date

The concept of inter-disciplinary interoperability and the need for it in managing societal issues is central to the addressing the challenges of sustainability research identified by ICSU. With this in mind, EuroGEOSS was launched on May 1st, 2009 for a three year period with the aim to demonstrate to the scientific community and society the added value of making existing earth observing systems and applications interoperable and used within the GEOSS and INSPIRE frameworks. The project builds an IOC in the three strategic areas of Drought, Forestry, and Biodiversity and undertakes the research necessary to develop this further into an Advanced Operating Capability (AOC) that provides access not just to data but also to analytical models

[5] http://www.eurogeoss.eu/

made understandable and useable by scientists from different disciplinary domains. The achievement of this AOC requires research in advanced modelling from multi-scale heterogeneous data sources, expressing models as workflows of geo-processing components reusable by other communities, and ability to use natural language to interface with the models.

The extension of INSPIRE and GEOSS components with concepts emerging in Web 2.0 communities with respect to user interaction and resource discovery, also supports the increased dialogue between science and society, which is crucial for building consensus on the collective action necessary to address global environmental challenges.

EuroGEOSS has completed the first half of its activities. During these first 18 months of the project, the key objectives were:

1. Achieving an IOC, i.e. the development of the services necessary to make it possible to discover view, and access the information resources made available by spatial data infrastructures (SDIs) available and developed by the partners of the project in the thematic areas of biodiversity, drought, and forestry.
2. Registering these resources as GEOSS components.
3. Developing the framework for assessing the added value of the project and of GEOSS to the communities of users.

All of these objectives have been achieved: the IOC in the fields of biodiversity, drought, and forestry has been established, it has been registered with GEOSS, and a multi-layered framework of surveys and models to assess the longitudinal impact of the project and the benefits of GEOSS have been put in place.

The Forestry IOC has been achieved giving priority to the development of federated metadata[6] catalogues and a map viewer, which are then integrated into the EuroGEOSS brokering framework. These priorities were expressed in an analysis of forestry users' requirements [2]. The IOC Metadata Catalogue was developed based on the open source package GeoNetwork v2.4.3 and populated with spatial and non-spatial metadata from the European Forest Data Centre at JRC. Metadata adjustments have been made to fit Dublin Core[7] elements and ensure compliance with INSPIRE and relevant ISO 19115 [9], ISO 19119 [10], and ISO 19139 [11] standards. The Metadata Catalogue functionalities and interface have been adjusted to meet the specific forestry theme requirements. As a result, the IOC Catalogue provides search, discovery and preview facilities of spatial and non-spatial metadata. The catalogue successfully harvests metadata from national and local forestry catalogues, such as those of the national Spanish spatial data infrastructure (IDEE), and is federated in the EuroGEOSS brokering framework so that its resources are globally accessible and viewable by the GEOSS community.

The Biodiversity IOC has been achieved based on the analysis of user requirements [15] by developing a series of metadata catalogues and services at the partners' institutions, and integrating them into the EuroGEOSS brokering framework. A key

[6] Metadata is a description of an information resource, including key elements such as what it is, who is responsible for it, where can it be found, and how it can be accessed.

[7] `http://dublincore.org/`

milestone has been the development of the metadata catalogue for the Global Biodiversity Information Facility (GBIF[8]) with a specialized profile using the Ecological Metadata Language to better support community needs, especially for species names datasets and natural history collections, and for multiple natural languages. A metadata sharing service has been established based on the Open Archive Initiative, harvesting metadata form the participating GBIF catalogues and integrating them into the EuroGEOSS brokering framework.

In parallel to this and related developments at other partners' institutions, significant work has taken place to develop a Digital Observatory for Protected Areas (DOPA[9]) a facility with initial focus on Africa but with a global reach as part of the GEOBON observation network [6]. DOPA will be developed in an iterative way, starting with an information system capable of visualizing and interacting through a single graphical user interface with key datasets hosted by the partners, namely boundaries of protected areas (United Nations Environment Programme - World Conservation Monitoring Centre, UNEP-WCMC), species occurrences (GBIF) and maps of bird distributions (Birdlife International and Royal Society for the Protection of Birds, RSPB). During the execution of the EuroGEOSS project, these developments will become more and more web-based, allowing the integration of information made available in the other thematic areas. The initial phase of the project has focused on the setting up of a prototype of DOPA that includes a specialized database, an advanced web client, and the preparation of unique datasets regarding bird distributions that will become available in the form of species occurrences via GBIF and in the form of species distribution maps directly through the DOPA.

The Drought IOC has been achieved by developing a series of web services to discover, view, and access drought data providers at the European level (EC Joint Research Centre, JRC), regional level (Observatory for South East Europe), and national/regional levels (Spanish Drought Observatory, and observatory for the Ebro river basin). The goal of connecting drought data providers on the three scale levels (continental, national/international, regional/local) was one of the key priorities expressed by users [7], and its achievement is an important proof of concept of a nested multi-scale system of systems. All the partners have in place an infrastructure for providing web map services (Open Geospatial Consortium - Web Map Server, OGC WMS) [14] and update their services regularly. Some partners (EC-JRC and University of Lubjana) provide also web map services of time series (WMS-T) for accessing data sets of a chosen date or period.

The integration of services from different partners in a common viewer, i.e. the map viewer of the European Drought Observatory (EDO), allows the linkage to services from the other thematic areas (e.g. forest) and opens new options for drought data analysis. These options will be further explored in the second half of the project. In addition to the European perspective, an interoperable EDO contributes to a future Global Drought Early Warning System under consideration by the World Meteorological Organization (WMO), and GEO/GEOSS. To this end, a prototype Global Drought Monitor has been established as a first building block of the Global Drought Early Warning System in partnership with the North American GEO/GEOSS

[8] http://www.gbif.org/

[9] http://dopa.jrc.ec.europa.eu/

community, the U.S. National Integrated Drought Information System (NIDIS) and the Princeton African Drought Monitor prototype. A first demonstration pilot of such Global Drought Monitoring System has been achieved and was demonstrated at the GEO Beijing Summit in November 2010[10].

Central to the inter-disciplinary IOC is the EuroGEOSS discovery broker, which is a component capable of reading and mediating among the many standards and specifications used by different scientific communities. By building bridges among the practices of these communities, the broker makes it possible to search, and discover the resources available from heterogeneous sources. During this initial phase, the EuroGEOSS discovery broker gives access to over 400 datasets and 26 services, including multiple catalogue services in the three thematic areas. By registering the broker as a GEOSS component, all of the thematic resources of the project are also accessible to the global research community. The following Section discusses briefly the key achievements and challenges of the brokering approach developed for EuroGEOSS.

3 The Brokering Approach

The EuroGEOSS inter-disciplinary IOC was built on the comparative analysis of the thematic user requirements [17] and is developed applying several of the principles/ requirements that characterize the System of Systems (SoS) approach and the Internet of Services (IoS) philosophy:

1. Keep the existing capacities as autonomous as possible by interconnecting and mediating standard and non-standard capacities.
2. Supplement but not supplant systems mandates and governance arrangements.
3. Assure a low entry barrier for both resource users and producers
4. Be flexible enough to accommodate existing and future information systems as well.
5. Build incrementally on existing infrastructures (information systems) and incorporate heterogeneous resources by introducing distribution and mediation functionalities to interconnect heterogeneous resources.
6. Specify interoperability arrangements focusing on the composability of inter-disciplinary concepts rather than just the technical interoperability of systems.

The key features of the EuroGEOSS inter-disciplinary IOC are the brokering and mediation capabilities that allow discovering and accessing autonomous and heterogeneous resources from the three thematic domains of the project. This is achieved by applying a *brokering* approach. This approach extends the traditional SOA archetype introducing an "expert" component: the Broker. It provides the necessary mediation and distribution functionalities to: *(i)* allow service consumer to

[10] See `http://www.ogcnetwork.net/pub/ogcnetwork/GEOSS/AIP3/pages/Demo.html` both Drought European and Drought Global.

bind to heterogeneous service providers in a transparent way, and *(ii)* to interact with them using a single and well-known end point. Such a solution addresses some of the shortcomings characterizing the present SOA implementations – like, semantic heterogeneity in resource descriptions and standard proliferation – which jeopardize the development of complex, large, and heterogeneous infrastructures, like GEOSS. Demonstrating the added value of this brokering approach is therefore one of the main contributions of EuroGEOSS to the development of GEOSS and the IoS.

The EuroGEOSS Discovery Broker provides the IOC with harmonized discovery functionalities by mediating and distributing user queries against tens of services presently registered in the EuroGEOSS capability – several of them are catalogs or inventory servers that propagate the query to many other resources. The key feature of the Discovery Broker is that it makes it possible for users to select among a list of well-adopted SOA and emerging Web 2.0 discovery interfaces, and easily utilize them. This list includes the service interfaces that comply with INSPIRE and/or OGC, service interfaces which are specific to the three thematic areas, and service interfaces which are well-used by other communities (e.g. Thematic Realtime Environmental Distributed Data Services, THREDDS[11] and Open-source Project for a Network Data Access Protocol, OPeNDAP[12]) or projects (e.g. Ground European Network for Earth Science Interoperations - Digital Repositories, GENESIS-DR[13] and SeaDataNet[14] [13]). Building these bridges to different communities makes it possible to serve the inter-disciplinary needs of scientific research without assuming that everyone will converge on one selected standard.

Figure 2 depicts the role played by the Discovery Broker in bringing together the capabilities provided by the three thematic areas and those shared by other Communities – e.g. Climatology, Meteorology, Oceanography and Hydrology. A partial list of the supported service interfaces is showed. The Discovery Broker is based on the GI-cat technology [12].

In order to facilitate inter-disciplinary data access, a specific brokering component has been introduced as part of the EuroGEOSS IOC. The EuroGEOSS Access Broker makes it possible for users to access and use the datasets resulting from their queries which are returned to them based on a common grid environment they have, previously, specified by selecting the following common features: Coordinate Reference System (CRS), spatial resolution, spatial extent (e.g. subset), data encoding format.

In keeping with the SoS principles, the EuroGEOSS Data Access Broker carries out this task by supplementing, but not supplanting, the access services providing the requested datasets. That is achieved by brokering the necessary transformation requests (those that the access services are not able to accomplish) to external processing services. Following the IoS and Web 2.0 principles, the broker publishes web applications allowing users to: *(i)* select a default business logic (i.e. algorithms) implementing dedicate processing like CRS transformation and space resolution resampling; *(ii)* upload their own business logic (i.e. processing schemes) and set it as

[11] `http://www.unidata.ucar.edu/projects/THREDDS/`

[12] `http://www.opendap.org/`

[13] `http://portal.genesi-dr.eu/`

[14] `http://www.seadatanet.org/`

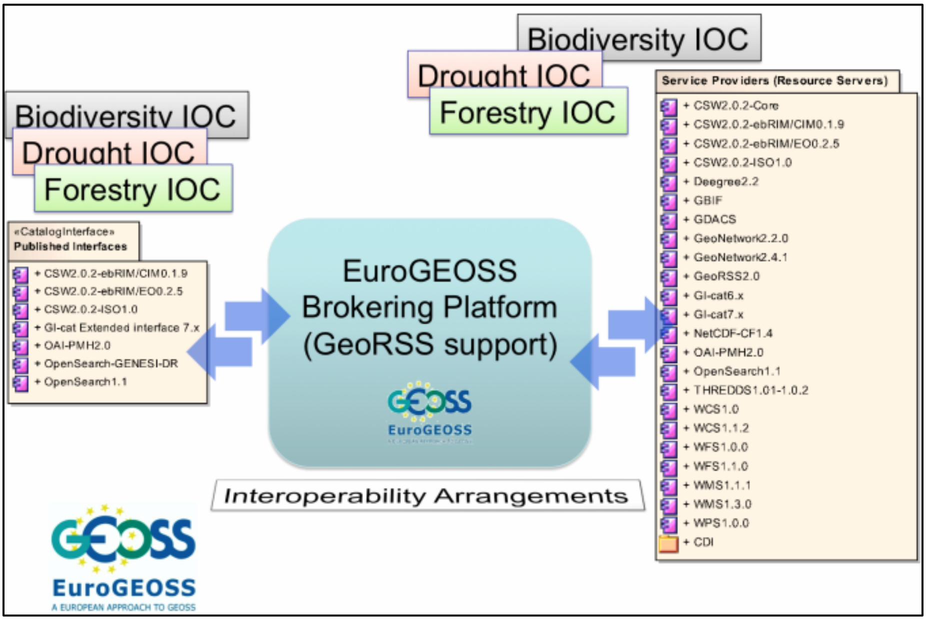


Fig. 2. Broker supporting multiple practices

default; *(iii)* select the order of the processing steps. The EuroGEOSS Data Access Broker also publishes an interface which realizes the INSPIRE transformation service abstract specification [8].

4 Enabling Semantic Inter-disciplinary Interoperability

As previous sections illustrated, each EuroGEOSS thematic area adopts a typical SDI framework which supports a variety of geographic data set types and geographical services. As experimented in the first phase of the project, each thematic area uses its internal vocabulary to annotate its resources. In order to enable the semantic interoperability among datasets and services that are provided by these heterogeneous thematic domains, a consensus on the different categorization is needed.

Usually, SDI frameworks do not provide any tool to enable semantic interoperability between different providers. For this reason, the EuroGEOSS project has developed, in collaboration with FP7 GENESIS[15] project, a solution to approach the semantic inter-disciplinary interoperability among the three EuroGEOSS thematic areas. The solution is based on three main components, which will be illustrated in the following subsections: (i) the annotation of resources made available by SDIs by referring to thesauri encoded in the *Simple Knowledge Organisation System* (SKOS) formalism, (ii) the provision of a matching tool (SKOSMatcher) [3] for harmonizing

[15] http://www.genesis-fp7.eu/

the thesauri that independent organisations may have adopted for the annotation of resources and (iii) a semantic Discovery Augmentation Component (DAC) [16] which harnesses the Discovery Broker capacity.

4.1 Organising Knowledge for SDIs

Discovering annotation of resources provided by the three EuroGEOSS thematic areas requires the adoption of new paradigms which are not available from the traditional Web (that is, the one made of HTML pages). The so-called "Web of Data", constituted by machine-accessible information, enables efficient search and retrieval of resources. The *Resource Description Framework* (RDF)[16] mechanism provides the building blocks to represent such machine-accessible information through a series of triples (subject-predicate-object). To associate specific semantic with data, RDF is extended by formalisms such as the *RDF Schema* (RDFS)[17] and the *Web Ontology Language* (OWL)[18].

In the following of this work, we will leverage RDF data structures for expressing SDI-related thesauri. These structures are based on an OWL ontology that was designed to provide a lightweight set of primitives for expressing *Knowledge Organisation Systems* (KOS) [1] that is, thesauri and classification schemata. This formalism is called *Simple Knowledge Organisation System* (SKOS)[19] which is general enough to express thesauri with no particular internal structure beside collections and concepts. Many general-purpose thesauri exist for annotating SDI-related resources. As an example, the EUROVOC[20] and AGROVOC[21] thesauri are provided as pure SKOS data sources. Instead the GEMET Thesaurus[22] and the INSPIRE Registry[23] [4] have been encoded according to an extension to the aforementioned schema in order to differentiate between orthogonal categorisations of terms, such as *groups* and *themes*. Thematic thesauri are also emerging in order to accommodate the specific *lingo* that specific thematic domains may use; for an example of these, you may browse the terms identified by the drought partners of EuroGEOSS (Drought vocabulary)[24], the Water ontology developed for the GEOSS AIP-3 "Semantics and Ontology Scenario"[25], and by the Spatial Information Service Stack Vocabulary hosted by the Australian CSIRO[26].

EuroGEOSS thesauri are currently accessible through a SPARQL[27] endpoint. SPARQL is a query language similar those used for relational data bases, but

[16] http://www.w3.org/TR/REC-rdf-syntax/

[17] http://www.w3.org/TR/rdf-schema/

[18] http://www.w3.org/TR/owl-features/

[19] http://www.w3.org/2004/02/skos/

[20] http://europa.eu/eurovoc/

[21] http://aims.fao.org/website/AGROVOC-Thesaurus

[22] http://www.eionet.europa.eu/gemet

[23] http://inspire-registry.jrc.ec.europa.eu/

[24] http://eurogeoss.unizar.es/home/

[25] http://www.ogcnetwork.net/pub/ogcnetwork/GEOSS/AIP3/pages/AIP-3_ER.html

[26] http://auscope-services.arrc.csiro.au/vocab-service

[27] http://www.w3.org/TR/rdf-sparql-query/

optimized for the querying of RDF triple stores, which are databases optimized for the storage and retrieval of RDF data. The only draw back of the general http SPARQL endpoint is that it can not be used to store information. In order to have a writable connection to store new relations between thesauri, aside the http endpoints, also SESAME[28] triple stores are supported using the SESAME-API.

4.2 Matching SKOS Vocabularies

The purpose of relating independent thesauri to each other is manifold: on the one hand, relations allow the user to navigate across distinct domains enabling expressive browsing functionalities; on the other hand, thesauri can more efficiently be used for describing and discovering spatial data and services among different disciplines. In fact, relating terms from distinct thesauri creates richer structural information that can be used for query expansion either with terms of a single thesaurus or multiple thesauri; see [5] for an advanced implementation scenario involving thesauri. This technique allows for narrowing or broadening the number of results returned by a query by referring to, respectively, more specific and more general terms in the hierarchy of terms either in a thesaurus or among multiple thesauri.

Despite the availability of thesauri in the SKOS format, there is no widely acknowledged application for aligning them. SKOS data structures that are made available according to the Linked Data[29] paradigm may take advantage of browsing tools that can be directly integrated into web browsers; nevertheless, they do not provide facilities for creating relations among the entities that are browsed. On the other hand, domain experts who may have the expertise to perform the alignment task might not be familiar with ontology editing tools, such as Protégé[30]; they may not even be knowledgeable of the SKOS and RDF data formats. Consequently, a tool to support them for performing this task without having to bother about the underlying data structures was needed.

SKOSMatcher is a web application, developed in collaboration with the FP7 GENESIS project, that allows for browsing and aligning thesauri available through generic HTTP/HTTPS SPARQL endpoints. The results of thesauri alignment by the domain expert is a set of triples linking terms in either an individual thesaurus or two independent thesauri. The new relations can be stored in any arbitrary instance of the Sesame RDF triple store[31] (since SPARQL does not provide repository update functionalities). The main information that is produced by creating relations between thesauri is, of course, the relations themselves; this is already sufficient to many applications (e.g., to perform query expansion and to provide correspondences between concepts of different disciplines). On the other hand, in order to ground governance and evolution of a relation set, it is necessary to collect more information and to associate metadata with the newly created relations (e.g., the rationale for a given relation, its creator, etc.).

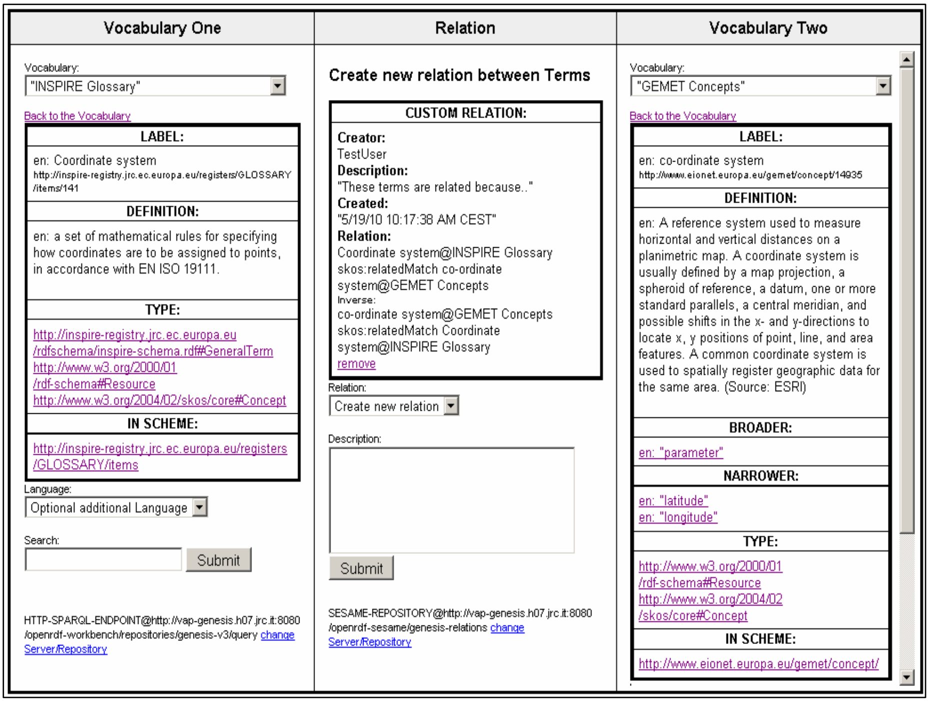
[28] http://www.openrdf.org/

[29] http://linkeddata.org/

[30] http://protege.stanford.edu/

[31] http://www.openrdf.org/

The SKOSMatcher has two modes of operation. It can be set up as a simple browser for SKOS thesauri or as a tool to browse two thesauri at the same time and create semantic relations between terms from each of them. In the latter configuration, two instances of the browser are combined with an additional pane to display the existing custom relations and to create new ones. Figure 3 is showing the details of a newly created relation between, e.g., the term “Coordinate system” from the INSPIRE Glossary (left hand side) and the term “co-ordinate system” from GEMET (right hand side). In the middle pane the details for a new relation are specified.



Vocabulary One

Vocabulary: "INSPIRE Glossary"

Back to the Vocabulary

LABEL:
en: Coordinate system
http://inspire-registry.jrc.ec.europa.eu/registers/GLOSSARY/items/141

DEFINITION:
en: a set of mathematical rules for specifying how coordinates are to be assigned to points, in accordance with EN ISO 19111.

TYPE:
http://inspire-registry.jrc.ec.europa.eu/rdfschema/inspire-schema.rdf#GeneralTerm
http://www.w3.org/2000/01/rdf-schema#Resource
http://www.w3.org/2004/02/skos/core#Concept

IN SCHEME:
http://inspire-registry.jrc.ec.europa.eu/registers/GLOSSARY/items

Language: Optional additional Language

Search:

HTTP-SPARQL-ENDPOINT@http://vap-genesis.h07.jrc.it:8080/openrdf-workbench/repositories/genesis-v3/query change Server/Repository

Relation

Create new relation between Terms

CUSTOM RELATION:
Creator:
TestUser
Description:
"These terms are related because.."
Created:
"5/19/10 10:17:38 AM CEST"
Relation:
Coordinate system@INSPIRE Glossary skos:relatedMatch co-ordinate system@GEMET Concepts
Inverse:
co-ordinate system@GEMET Concepts skos:relatedMatch Coordinate system@INSPIRE Glossary
remove

Relation: Create new relation

Description:

SESAME-REPOSITORY@http://vap-genesis.h07.jrc.it:8080/openrdf-sesame/genesis-relations change Server/Repository

Vocabulary Two

Vocabulary: "GEMET Concepts"

Back to the Vocabulary

LABEL:
en: co-ordinate system
http://www.eionet.europa.eu/gemet/concept/14935

DEFINITION:
en: A reference system used to measure horizontal and vertical distances on a planimetric map. A coordinate system is usually defined by a map projection, a spheroid of reference, a datum, one or more standard parallels, a central meridian, and possible shifts in the x- and y-directions to locate x, y positions of point, line, and area features. A common coordinate system is used to spatially register geographic data for the same area. (Source: ESRI)

BROADER:
en: "parameter"

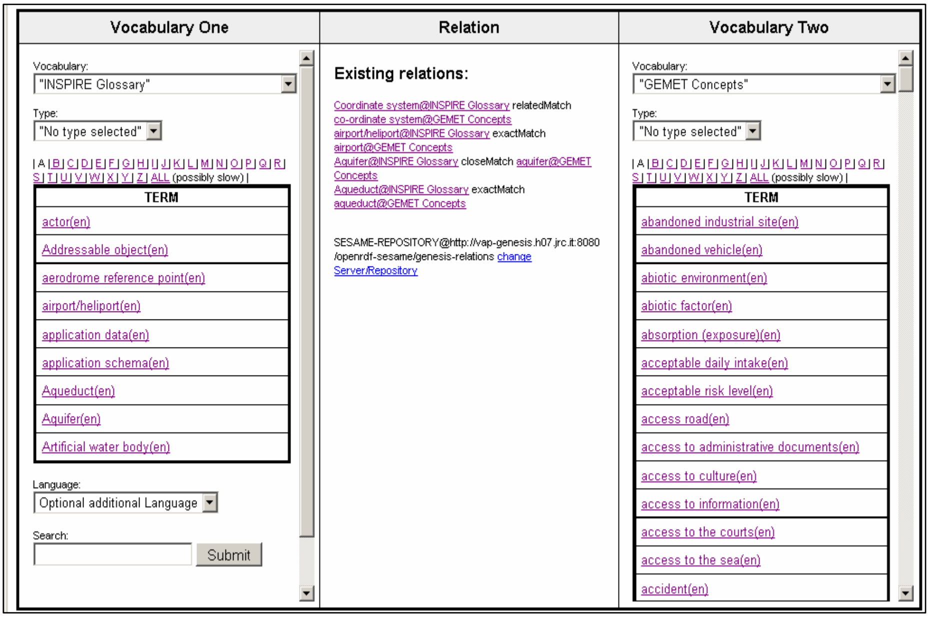
NARROWER:
en: "latitude"
en: "longitude"

TYPE:
http://www.w3.org/2000/01/rdf-schema#Resource
http://www.w3.org/2004/02/skos/core#Concept

IN SCHEME:
http://www.eionet.europa.eu/gemet/concept/

Fig. 3. The details of a newly created relation

As soon as there are existing relations, the middle pane can also be used to browse them. After selecting a thesaurus in either of the browsing panes, the existing connections between the thesaurus and any other thesaurus that is available through the other endpoint are displayed. With the selection of a second thesaurus in the other pane, the list is narrowed down to the connections between the two selected thesauri. Relations contained in each of the thesauri that are displayed are not displayed in the middle pane; instead, they can be found in the side panes where they provide navigation functionalities to browse the thesauri. Figure 4 is showing a list of sample relations that has been created between the GEMET Thesaurus and the INSPIRE Glossary.



Vocabulary One	Relation	Vocabulary Two
Vocabulary: "INSPIRE Glossary" Type: "No type selected"	**Existing relations:** Coordinate system@INSPIRE Glossary relatedMatch co-ordinate system@GEMET Concepts airport/heliport@INSPIRE Glossary exactMatch airport@GEMET Concepts Aquifer@INSPIRE Glossary closeMatch aquifer@GEMET Concepts Aqueduct@INSPIRE Glossary exactMatch aqueduct@GEMET Concepts SESAME-REPOSITORY@http://vap-genesis.h07.jrc.it:8080 /openrdf-sesame/genesis-relations change Server/Repository	Vocabulary: "GEMET Concepts" Type: "No type selected"

| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R |
S | T | U | V | W | X | Y | Z | ALL (possibly slow) |

TERM
actor(en)
Addressable object(en)
aerodrome reference point(en)
airport/heliport(en)
application data(en)
application schema(en)
Aqueduct(en)
Aquifer(en)
Artificial water body(en)

Language: Optional additional Language

Search:

| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R |
S | T | U | V | W | X | Y | Z | ALL (possibly slow) |

TERM
abandoned industrial site(en)
abandoned vehicle(en)
abiotic environment(en)
abiotic factor(en)
absorption (exposure)(en)
acceptable daily intake(en)
acceptable risk level(en)
access road(en)
access to administrative documents(en)
access to culture(en)
access to information(en)
access to the courts(en)
access to the sea(en)
accident(en)

Fig. 4. Upon selection of thesauri the relations linking them are shown

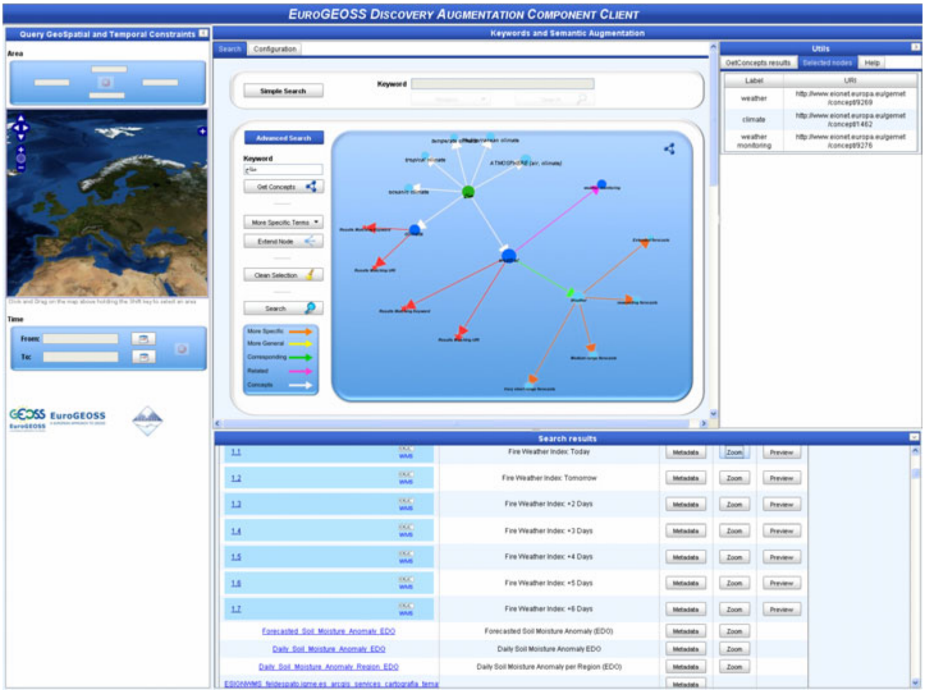
4.3 Augmenting Semantic Discovering Capabilities

The EuroGEOSS semantic Discovery Augmentation Component (DAC) implements a “third-party discovery augmentation approach”: enhancing discovery capabilities of infrastructures by developing new components that leverage on existing systems and resources to automatically enrich available geospatial resource description with semantic meta-information. DAC provides users with semantics enabled query capabilities – contributing to bridge a gap which is important for inter-disciplinary SOA infrastructures.

The EuroGEOSS DAC federates both multilingual controlled vocabularies providing semantics (i.e. Simple Knowledge Organization System, SKOS, repositories) and ISO-compliant geospatial catalogue services. In fact, the EuroGEOSS DAC is able to use existing discovery (e.g. catalogs and discovery brokers) and semantic services (e.g. controlled vocabularies, ontologies, and gazetteers). As illustrated in section 3, in our case, ISO compliant geospatial catalog services are provided by the catalogs of the three thematic areas connected to the EuroGEOSS broker and developed during the IOC phase. Multilingual controlled vocabularies are provided by the EuroGEOSS SPARQL endpoint and managed through the SKOSMatcher tool, as illustrated in 4.1 and 4.2.

The DAC can be queried using common geospatial constraints (i.e. what, where, when, etc.) and currently, two different augmented discovery styles are supported: *(i)* automatic query expansion; *(ii)* user-assisted query expansion. Both styles execute a look-up of the query string inserted by the user in the thesauri that are provided by the GENESIS vocabulary service, allowing to match language-dependant search patterns against language-neutral URIs. The result of this process is a list of URIs corresponding to terms. In the first discovery mode, these URIs are first enriched with those of terms that are semantically related to those matching the user-defined query string. The specific axes (e.g. more specific terms, related terms, etc.) along which the new terms are retrieved can be configured at setup-time. The resulting set of URIs is then translated back into a number of different languages, also customizable at setup-time, and the resulting text patters are used for executing multiple queries to catalogs.

More interesting is the second approach, shown in Figure 5, allowing the user to freely navigate the thesauri and define a custom set of terms according to which the search shall be carried out.



Label	URI
weather	http://www.eionet.europa.eu/gemet/concept/9269
climate	http://www.eionet.europa.eu/gemet/concept/1462
weather monitoring	http://www.eionet.europa.eu/gemet/concept/9276

1.1	OGC WMS	Fire Weather Index: Today	Metadata	Zoom	Preview
1.2	OGC WMS	Fire Weather Index: Tomorrow	Metadata	Zoom	Preview
1.3	OGC WMS	Fire Weather Index: +2 Days	Metadata	Zoom	Preview
1.4	OGC WMS	Fire Weather Index: +3 Days	Metadata	Zoom	Preview
1.5	OGC WMS	Fire Weather Index: +4 Days	Metadata	Zoom	Preview
1.6	OGC WMS	Fire Weather Index: +5 Days	Metadata	Zoom	Preview
1.7	OGC WMS	Fire Weather Index: +6 Days	Metadata	Zoom	Preview
Forecasted Soil Moisture Anomaly EDO		Forecasted Soil Moisture Anomaly (EDO)	Metadata	Zoom	
Daily Soil Moisture Anomaly EDO		Daily Soil Moisture Anomaly EDO	Metadata	Zoom	
Daily Soil Moisture Anomaly Region EDO		Daily Soil Moisture Anomaly per Region (EDO)	Metadata	Zoom	
ESIGNWMS feldespato.igme.es arcgis services cartografia tema			Metadata		

Fig. 5. Broker supporting multiple practices

In the Figure, the Arabian search string " مناخ " (“climate”) is found in the definitions of seven terms from GEMET. Starting from these, the user has extended term “weather”, moving from the original thesaurus to the corresponding term from GEOSS Societal Benefit Areas (SBAs). This term is then expanded by searching for more specific terms. The user can now select the terms that are matching its needs best and execute the actual query against catalogue services. Again, the resulting set of URIs is translated back into different languages and queries to catalogs are executed. The EuroGEOSS brokering platform, propagates the query to the federated catalogs, collects and provides the resulting records to the DAC which, in turn, shows the results (see Figure 5, bottom part of the DAC GUI).

5 Discussion and Next Steps

There exist clear challenges on using and integrating inter-disciplinary resources to develop cross-disciplinary applications. They include:

- High Entry Barrier: users need to “learn” and develop many (sometimes, immature) information technologies.
- Limited functionalities: international community has mainly focused on discovery functionality implementation; while, cross-domain evaluation, transfer and use functionalities are still lacking.
- Limited semantic interoperability: interoperability for heterogeneous disciplinary resources and different domain semantics are still main issues.
- Limited sustainability: as for scalability, a flat approach to interconnect resources is not sustainable in presence of hundreds of thousands of (heterogeneous) entries and hundreds of registered standards; as for flexibility, future systems and specifications must be easily added, as well.

EuroGEOSS experimented a brokering framework to address these challenges. In fact, this solution can provide a homogeneous discovery, evaluation, and access framework to heterogeneous resources in a seamless way for users –lowering the entry barrier. It is able to implement conceptual composability (not just technical interoperability) allowing a major flexibility and scalability. It can enable semantic query by making use of existing semantic engines, developed by the diverse communities, preserving their autonomy and replacing them, if necessary.

This is achieved by extending the SOA approach and advancing it through the use of “expert” components, such as, for example the EuroGEOSS broker, knowledge management systems and advanced semantic discovery tools[32]. During the next 18 months the project will build its AOC, so that it is possible to access and use not just data across multiple thematic areas but also models and analytical process expressed in workflows and implemented through web-based chains of services. The main expected impact of this development is to make the EuroGEOSS resources accessible and usable not only from specialists in the individual fields, but also from scientists from multiple disciplines that will be able to have a clear picture of how the resources available can be used to address specific questions and how they may be adapted for their specific needs. In addition, the work already started in the project on natural language interfaces, and semantic inter-disciplinary interoperability which was presented in the last part of this paper. Also, lessons to be learned from Web 2.0 social networks offer the opportunity to expand the use of the EuroGEOSS infrastructure to a much wider audiences that transcend scientific disciplines.

References

1. Bechhofer, S., Goble, C.A.: Thesaurus construction through knowledge representation. Data Knowl. Eng. 37, 25–45 (2001)
2. Figueiredo, C., Gaigalas, G., Achard, F., Iglesias, J.M.: Report on User Requirements for the EuroGEOSS Forestry Operating Capacity. EuroGEOSS D3.1, Technical report (2009), `http://www.eurogeoss.eu/Documents/EuroGEOSS_D3_1.pdf`

[32] See Also: GeoS2011 ‘An Approach to the Management of Multiple Aligned Multilingual Ontologies for a Geospatial Earth Observation System’.

3. Fugazza, C., Dupke, S., Vaccari, L.: Matching SKOS Thesauri for Spatial Data Infrastructures. In: Sánchez-Alonso, S., Athanasiadis, I.N. (eds.) MTSR 2010. CCIS, vol. 108, pp. 211–221. Springer, Heidelberg (2010)
4. Fugazza, C., Dupke, S., Schade, S.: Semantics-Aware Thesauri Management for Advanced Spatial Data Retrieval. Submitted to the special issue of the World Wide Web - Internet and Web Information Systems Journal, Springer Querying the Data Web. Novel techniques for querying structured data on the web (2010)
5. Fugazza, C., Chinosi, M., Luraschi, G.: Ontology-based governance of INSPIRE metadata: An implementation scenario. In: INSPIRE Conference, Krakòw, Poland (2010)
6. GEO: GEO Biodiversity Observation Network Concept Document, GEO-V document 20, Geneva, Switzerland, Technical report (2008), http://www.earthobservations.org/documents/cop/bi_geobon/200811_geobon_concept_document.pdf
7. Hofer, B., Niemeyer, S., Ceglar, A., et al.: Report on User Requirements for the EuroGEOSS Drought Operating Capacity. EuroGEOSS D5.1, Technical report (2010), http://www.eurogeoss.eu/Documents/EuroGEOSS_D_5_1.pdf
8. Howard, M., Payne, S., Sunderland, R.: Technical Guidance for the INSPIRE Schema Transformation Network Service (2010), http://inspire.jrc.ec.europa.eu/documents/Network_Services/JRC_INSPIRE-TransformService_TG_v3-0.pdf
9. ISO: Geographic Information – Metadata, ISO/IEC_19115:2003 (2003)
10. ISO: Geographic Information – Services, ISO_19119:2005 (2005)
11. ISO: Geographic Information – Metadata – XML Schema Implementation, ISO/TS_19139:2007 (2007)
12. Nativi, S., Bigagli, L.: Discovery, Mediation, and Access Services for Earth Observation Data, Selected Topics in Applied Earth Observations and Remote Sensing. IEEE Journal 2(4), 233–240 (2009)
13. Nativi, S., Mazzetti, M., Santoro, M., Boldrini, E., Manzella, G.M.R., Schaap, D.M.A.: CDI/THREDDS interoperability in the SeaDataNet framework. Adv. Geosci. 28, 17–27 (2010), doi:10.5194/adgeo-28-17-2010
14. OGC: Web Map Server (WMS) Implementation Specification ver. 1.3.0 (2006)
15. O' Tuama, E., Dubois., G., Cottam, A., May, I., Fisher., I.: Report on User Requirements for the EuroGEOSS Biodiversity Operating capacity. EuroGEOSS D4.1, Technical report (2009), http://www.eurogeoss.eu/Documents/EuroGEOSS_D4_1.pdf (last date accessed 1/2011)
16. Santoro, M., Mazzetti, P., Fugazza, C., Nativi, S., Craglia, M.: Semantics Enabled Queries in EuroGEOSS: a Discovery Augmentation Approach, AGU (2010)
17. Vaccari, L., Nativi, S., Santoro, M.: Report on Requirements for Multidisciplinary Interoperability. EuroGEOSS D2.1.1 (2010), http://www.eurogeoss.eu/Documents/EuroGEOSS_D_2_1_1.pdf (last date accessed 1/2011)

Improving Geodatabase Semantic Querying Exploiting Ontologies

Miriam Baglioni[1,*], Maria Vittoria Masserotti[2], Chiara Renso[2],
and Laura Spinsanti[3,**]

[1] IIT, CNR, Via Moruzzi 1, 56100, Pisa, IT
baglioni@iit.cnr.it
[2] KDDLab, ISTI, CNR, Via Moruzzi 1, 56100, Pisa, IT
{masserotti,renso}@isti.cnr.it
[3] IES, JRC, European Commission, Ispra, IT
laura.spinsanti@jrc.ec.europa.eu

Abstract. Geospatial semantic querying to geographical databases has been recognized as an hot topic in GIS research. Most approaches propose to adopt an ontology as a knowledge representation structure on top of the database, representing the concepts the user can query. These concepts are typically directly mapped to database tables. In this paper we propose a methodology where the ontology is further exploited mapping axioms to spatial SQL queries. The main advantage of this approach is that semantic-rich geospatial queries can be abstractly represented in the ontology and automatically translated into spatial SQL queries.

Keywords: semantic geospatial query, ontologies, spatial database, spatial SQL.

1 Introduction

The exponential growth of positioning technologies coming from the widespread use of GPS personal devices, for example embedded in smart phones, tend to produce a huge amount of geographical referred data. This, in turn, calls for methods and techniques capable of managing and querying this large quantity of geo-referenced data. Moreover many non-expert users are becoming aware of the huge potentiality of geographic information in their everyday life. While data management recent developments in spatial and spatio-temporal databases converge towards some well-defined proposals [OraSpa, PostGIS, Guting05], the query capabilities in terms of semantically enhanced user query language have not produced so far any standardized approach. In this context, proposals to facilitate the access to geographical data to the non-expert user may range from advances graphical user interfaces and visual analytics to natural language processing. In the latter approaches having support for semantic geospatial queries is essential and usually these proposals employ an ontology [Gru08] as a knowledge representation level on top of a spatial database. The objective of the ontology layer is to explicitly represent the concepts the user can mention in the query.

[*] The work was carried out at Dep. Computer Science, Univ. of Pisa, 56100, Pisa, IT.
[**] The work was carried out at LBD Lab EPFL, EPFL-IC-LBD, CH-1015 Lausanne, CH.

C. Claramunt, S. Levashkin, and M. Bertolotto (Eds.): GeoS 2011, LNCS 6631, pp. 16–33, 2011.

In the last decade, ontologies have gained increasing interest in the GIS community [Mark06, Fonseca02, Spaccapietra04], because they can be used to create and exploit data standards as well as human computer interfaces and to solve heterogeneity/ interoperation problems. The relevant literature exploits ontologies to map data sources to explicit ontology concepts [Bishr08] or in geographical information retrieval techniques [Mata09]. The use of ontologies as a middle layer between the user and the database adds a conceptual and semantic level over the data, and allows the user to query the system on semantic concepts without having any specific information about the database at hand [Peuquet02]. Despite of the efforts to create sharable data, many Geographical Information Systems (GIS) evolve from numeric cartography integrating remote sensing and digital images, typically skipping any design and modeling phase. Therefore, quite often they lack both metadata and the conceptual schema, thus losing part of the semantic geographical information. An effect of having an ontology layer linked to a geographical database is to increase not only the number of (geographical) concepts the user can mention in the query - including all the ontology concepts that are abstractions of the database tables - but also the "abstraction level" in term of semantic definitions. To better clarify this point let us consider for example a database storing tables of urban objects such as schools, hospital, universities. Assume that thesetables are mapped to urban ontology concepts and subsumed in the ontology by a *Building* concept representing all the spatial objects in the domain which are buildings. This means allowing the user to mention the Building concept in the query abstracting away from the specific kind of building such as schools, hospitals etc.

Given this scenario the main contribution of the present work is to map to the database not only the ontology classes but also the ontology concepts defined by *axioms*, the ontology implicit formal concept definitions. For example, the concept *FloodRiskBuldings*, may define the class of buildings which are *inside* a flood risk area. When these axioms represent spatial relations (in a given form, see Section 4) we provide a method to automatically translate them into spatial database materialized views. The consequence is that the user may query the database not only in the stored tables, but also in complex spatial queries abstractly represented in the ontology and stored as materialized views.

The proposed methodology intends to exploit the already existing domain ontologies where concepts are implicitly defined by axioms, as a conceptual representation of spatial SQL queries. Hence, the semantics of the underlying geographical database is enhanced by explicitly representing information which is implicit in the data (such as the spatial relations) and which can be retrieved by specific spatial SQL queries. We automatically map these ontology axioms into appropriate spatial SQL queriesstored asmaterialized views in the database. We believe that natural language processing techniques may benefit from the introduced methodology since there is a need to have a formal definition of the concepts that may be mentioned in the query. This further increases the expressive power of the query system: in fact, it increases the number of concepts the user can query with respect to the classical ontology database mapping.

Paper is organized as follows. Section 2 presents the approach, while Section 3 shows in details the semantic enhancing process. Then, Section 4 presents the algorithm and Section 5 the related work. Finally, Section 6 draws some conclusions and future work.

2 The Overview of the Approach

In this section we give the overall view of the proposed approach, firstly introducing the main idea of the methodwith a simple example, then describing some background concepts and presenting the methodology.

The problem statement is given in an urban scenario describing hospitals, schools, and flood risk areas of the citystored in a spatial database, as depicted below:

Table 1. The Flood Risk example

HOSPITAL			
ID	Name	The_geom	Helicopter
231	Careggi	POLYGON(10.589,47.779,…,WGS84)	Y
254	S.Maria	POLYGON(10.579,47.783,…,WGS84)	N

SCHOOL				
ID	Name	Classes	Category	The_geom
45	L.daVinci	52	Secondary	POINT(10.583,47.775,…,WGS84)
83	Mazzini	15	Primary	POINT(10.577,47.789,…,WGS84)

FLOOD RISK BASIN		
ID	Code	The_geom
26	AZ34	POLYGON(10.581,47.749,…,WGS84)
42	3SX9	POLYGON(10.579,47.782,…,WGS84)

Then consider the natural language query: *Which are the public buildings in a flood risk area*? The information to answer this query is actually stored in the spatial database, since we have information about hospitals and schools (which are public buildings) as well as the flood risk areas. However, public buildings have no correspondent database table where to issue a spatial SQL query. This means a semantic interpretation step has to be applied to the query: the user must query the HOSPITAL and SCHOOL tables and, from the resulting records, she/he has to select the ones which have an *intersect* spatial relationship with the flood risk basin of the city. Obviously, to express these queries the user needs to have the knowledge of both the database structure and the appropriate spatial operator to apply, depending on the type of the referred spatial objects.

The general objective of the present work is to introduce a methodology to support the automatic translation of spatial semantic information represented in an ontology into spatial SQL. The knowledge contained in the ontology can be of two kinds. First, knowledge is organized in a taxonomy of geographical places so that the more general concept are represented as top concepts, compared to the more specific concepts, correspondent to database tables, represented as “leaves” of the taxonomy. Secondly, we assume some concepts are defined as a resultof a spatial relation between two spatial objects and formalized as axioms. The central point of the work is to show that it is possible to map geographic places of interest, conceptually expressed by ontology axioms, to a query over the actual data stored in a geographical database. The advantage of using an ontology rely on the fact that the user can query the data without having the knowledge of the structure of the underlying database, or the spatial SQL syntax.

2.1 Background

Ontologies have certainly become a research topic in several disciplines, ranging from philosophy, geography, geomatics up to machine learning and artificial intelligence. The definition given by [Gru08] is used to define ontology as "a technical term denoting an artifact that is designed for a purpose, which is to enable the modeling of knowledge about some domain, real or imagined". Such ontologies determine what can be represented and what can be inferred about a given domain, using a specific formalism of concepts. Formally, an ontology is a 5-tuple *O:={C, R, HC, rel, A0}*, where *C* is a set of concepts, which represent the entities in the ontology domain; R is a set of relations defined among concepts; *HC* is a taxonomy or concept-hierarchy, which defines the *is_a* relations among concepts (HC(C1, C2) means that C1 is a sub-concept of C2), *rel:* $R \rightarrow C \times C$ is a function that specifies the relations on R. Finally, *A0* is the set of axioms expressed in a logical language, such as first order logic.

Ontology languages are formal languages used to construct ontologies. They allow the encoding of knowledge about specific domains and often include reasoning facilities that support the processing of that knowledge. Among all the ontology languages, we considered Web Ontology Language (OWL), that is a well known standard arisen from the Semantic Web and it is now a W3C recommendation [OWL]. An interesting feature of OWL is that it relies upon a family of languages known as Description Logics (DL) that provide a deductive inference system based on a formal well founded semantics [Baad03]. The basic components of DL are concepts (classes) and roles (properties), termed as TBox, and individuals (instances), termed as ABox. Concepts describe the common properties of a collection of individuals and roles are binary relations between concepts. Furthermore, a number of language constructs, such as intersection, union and role quantification, can be used to define new concepts and roles, by means of axioms.

2.2 Methodology

Our approach is based on the use of an ontology to represent the domain of interest and as a middle layer between the user and the data. The connection between the ontology and the database can be done by an explicit mapping between database tables and ontology concepts using a correspondence table. The mapping of concepts to database tables can be done in several ways, such as manually, or automatically extracting the ontology from the database [Baglioni07, Vital09]. The ontology – to – database mapping is deeply covered in the literature, and details of the different approaches can be found for example in [Barrasa04, An06, Li05].

Let us assume to have a geographical database containing the following tables: Hospital, School, Street, Square, PedestrianArea, CityCenter, River and FloodRisk Basin. And let us suppose to have the domain described by the ontology in Figure 1. As a matter of terminology, we refer to the following definitions to describe the possible ontology-to-database mappings:

- **Direct:** There is a one to one correspondence between an ontology concept and a database table. Considering the database tables listed above, and the ontology of Figure 1, there is direct correspondence for the ontology concepts Hospital, School, Street, Square, PedestrianArea, CityCenter, River and FloodRiskBasin.

- **Indirect:** The concept is an antecedent in the hierarchy (a super-class) of a concept that has a direct connection to the database. Considering the database tables listed above, and the ontology of Figure 1, it is possible to indirectly refer the PublicBuilding, District, Place, Building, UrbanObject and GeographicObjects concepts.

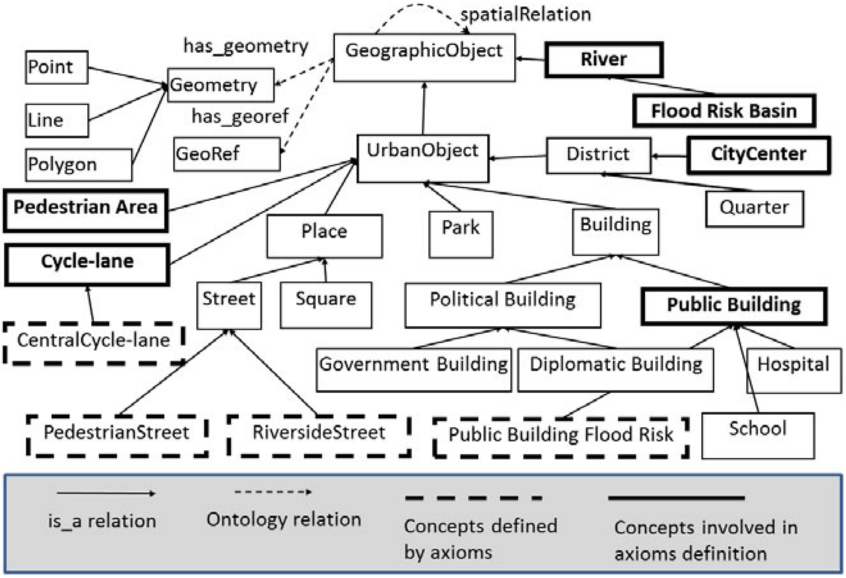


Fig. 1. The Urban Ontology

- **Implicit:** The concept is defined by an axiom and it is not directly connected to the database. In the example it is possible to implicitly refer the PublicBuilding FloodRisk, PedestrianStreet, and RiversideStreet concepts.

The main contribution of this paper is a methodology to automatically map implicit concepts to SQL queries. Axioms are expressed in OWL DL describing properties the entities must have to belong to the concept.

In the ontology in Figure1, boxes in dotted line represent the concepts defined by axioms. The form of axioms we are assuming here use a spatial relation between two geographical objects to define a new concept. For example, the dotted boxes in Figure 1 are defined as follows:

- *Public Building Flood Risk* be the Public buildings located *inside* the Flood Risk basin
- *PedestrianStreet* be a Street located (at least partially) inside the PedestrianArea
- *RiversideStreet*be a Street located along the River (within a distance of maximum 200m)
- *CentralCycle-lane* be the Cycle-lane that intersects the CityCenter.

These descriptions can be formalized in OWL DL axioms having the general form:

SpecializationConcept≡Concept1 and (spatialRelation some Concept2)

Again, consider the concept PublicBuildingFloodRisk that can be expressed combining PublicBuilding with FloodRiskBasin through the spatial property *intersect.* As an OWL axiom, this can be formalized as:

PublicBuildingFloodRisk≡PublicBuilding and (intersect some FloodRiskBasin).

A schema of the approach is illustrated in Figure 2, where the ontology is analyzed by an *Enhancing Module* in order to discover the axioms that can be mapped to SQL queries. The Enhancing Module, given a domain ontology and a geographical database, produces two outcomes: a number of SQL queries (stored in the DBMS as materialized views) which are derived by ontology axioms and a new DataBase Ontology (DBOntology). The DBOntology is obtained from the original domain ontology selecting only the concepts that have a direct, indirect and implicit connection to the database. Therefore, the DBOntology acts as a knowledge representation structure which is the basis of the query module, representing all the concepts that the user can mention in the query. The query module could be a complex natural language module (like in [Baglioni08, Baglioni09]) or a simpler interface depending on the application. The query module details are not reported in this paper.

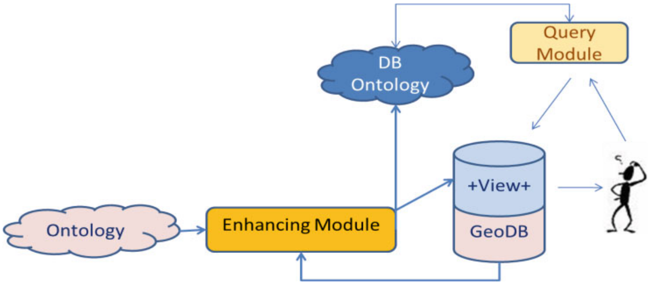


Fig. 2. The proposed methodology

An example of DBOntology resulting from the application of the Enhancing Module to the ontology of Figure 1 and the database tables listed above, is reported in Figure 3. Here, the white solid rectangles represent the classes of the ontology that define the spatial topological relations. The spatial relations considered here are inspired by the OpenGIS [OGIS] standard and are: *equals, disjoint, touches, within, overlaps, crosses, intersect, contains*. These relations are modeled in the ontology as sub-properties of *spatialRelation* (see Figure 1). The dark solid rectangles represent the classes of the ontology that have either a *direct* or an *indirect* mapping to the database. The white shattered rectangles represent concepts that have an *implicit* mapping to the database.

From Figure 3, we can notice that Public Building is a super-class of the two concepts Hospital and School, which, in turn, have a direct correspondence to database tables. The Public Building Flood Risk concept is defined by an axiom stating that a flood risk public building is a kind of PublicBuilding that *intersects* the

FloodRiskBasin. Since FloodRiskBasin has a direct map to a database table, the axiom related to the concept PublicBuildingFloodRisk can be translated to a database query. In the same way, Riverside Street and PedestrianStreet concepts can be mapped to spatial SQL queries. It is worth noticing that some concepts of the domain ontology have no mapping with the data (i.e Cycle-lane, DiplomaticBuilding, Park) therefore they will not appear in the final DBOntology. For the same reason the CentralCycle-lane concept, defined by an axiom, is excluded by the algorithm because it refers to the concept Cycle-lane that is not represented in the database.

As mentioned above, we base our approach upon OWL DL, a W3C standard arisen from research on the Semantic Web based on Description Logics [Baad03]. However, the kind of axioms that we consider have a specific simple form to ensure an easy translation into spatial SQL. Indeed, mapping OWL into SQL is a complex task object of research since many years and it is out of the scope of the paper presenting all these approaches (see for example [Acci05, Calvanese09]). The main difference among literature approaches and the present work is that we use OWL axioms only from a *syntactical* point of view in order to translate into the appropriate SQL query. We are not interested in the *reasoning* aspects of the ontology and axioms. The OWL fragment we consider here has not relations to the inference power of the reasoning engine.

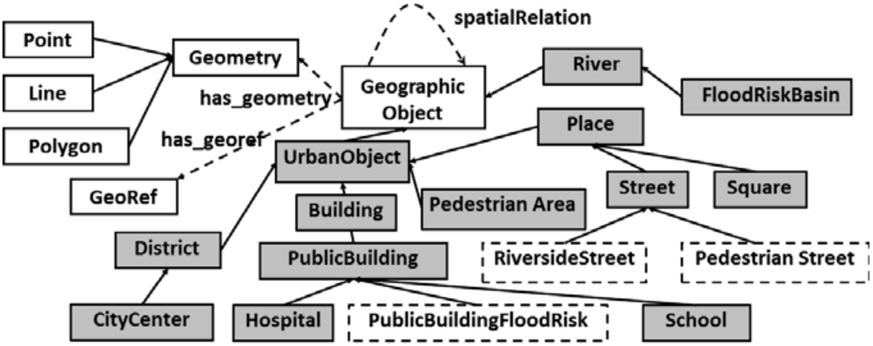


Fig. 3. The DBOntology where only the concepts mapped to the database are represented

More in general, we can say that allowed axioms define specialization classes by restricting the spatialRelation property (or its sub properties) with the "exists" operator (also called *some* in OWL). Indeed, this logical expression may be mapped to a SQL SELECT statement involving the spatial database topological relations. We explicitly refer to a spatial relation which can be immediately one-to-one mapped in a spatial operator.

In the next section we introduce the details of the semantic enhancing process with an exhaustive explanation of all the possible cases which can be found and processed.

3 The Semantic Enhancing Process

The Enhancing module is responsible for both the selection of the domain ontology classes that can be solved as an SQL query, the production of the corresponding

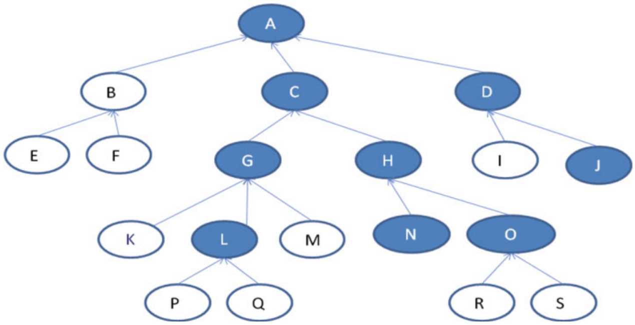


Fig. 4. An example of a generic ontology schema Onto

materialized views and contextually, the construction of the DBOntology. We illustrate the process by means of an example.

Consider a generic ontology schema, indicated as Onto, shown in Figure 4. Dark circles represent the ontology concepts *directly or indirectly* connected to the database, whereas white circles represent concepts *implicitly* connected to the database. Concepts that are not directly, indirectly, or implicitly mapped to database tables are not depicted in the figure. The enhancing process firstly considers all the concepts in Onto that have a *direct* connection to the database. They correspond to the set of concepts {J, L, N, O}. Among them, we select their direct children and sibling that have an *implicit* mapping, shown in Figure 4 in white color. Hence, the selection of the concept O will cause the selection of the nodes R and S since both are direct children of O. The concept L has both siblings and children and will cause the selection of the concepts {K, M, P, Q}. The concept N does not have any direct children, and the only sibling is already part of the dark concepts of Onto (direct and indirect mapped), so nothing else is selected. Therefore, the set of the concepts in Onto selected in this first step as potentially mapped to SQL queries is {I, K, M, P, Q, R, S}.

In the next step the enhancing process checks if these selected concepts can be associated to materialized views. To do this we must ensure that: (i) the axiom that defines the concept can be expressed by a spatial SQL query and (ii) the concepts defined by axiom are already in the DBOntology which has been built.

The DBOntology is built iteratively, and at the first step it is composed of all the dark concepts in Onto. The Enhancing Module checks the white concepts (implicit mapped), and when axioms are defined in terms of a spatial relation with a concept belonging to the DBOntology iteratively built so far, this new concept incrementally becomes part of the new DBOntology. This process is recursive and finite since the number of considered concepts is finite. This process results in the construction of the DBOntology that becomes the knowledge representation structure mapped to the spatial database where both direct, indirect and implicit connection to the database is explicitly represented.

We can define formally this process as follows.

Given a domain Generic Ontology O_G=(C_G,R_G,rel_G,H_G,A_G), the DBOntology O_{DB} =(C_{DB},R_{DB},rel_{DB},H_{DB},A_{DB}) is defined as follows:

- The set of ontology concepts is defined recursively starting from the concepts in C_G so

  $$C_{DB}^{0} = C_G$$

  $$C_{DB}^{n} = C_{DB}^{n-1} \cup \{c \in C_{DO} | \exists c' \in Leaf(C_G) \wedge (h_{DO}^{-1}(c) = h_{DO}^{-1}(c') \vee H_{DO}(c,c')) \wedge c' \neq c \wedge \exists a \in A_{DO}.c \in Conc(a) \wedge \forall c'' \in Conc(a).\ c'' \neq c \wedge c'' \in C_{DB}^{n-1}\}$$

  where
  - C_{DO} are the concepts in Onto,
  - $Leaf(C)=\{c\in C | \neg\exists c' \in C.H(c,c')\}$,
  - $h^{-1}:C \rightarrow C.\ h^{-1}(c)=c' \Leftrightarrow H(c',c)$
- $A_{DB} = \{a\in A_{DO} | \forall c\in Conc(a).\ c\in C_E \wedge \forall r\in relations(a).r\in SpatialRelation\}$

  where
  - A_{DO} is the set of axioms in Onto
  - $Conc(a)=\{c\in C|\ a\in A,\ \ c\in a\}$ is the set of concepts that define the axiom a
  - $relations(a)=\{r \in R|\ a\in A,\ r\in a\}$ is the set of relations that define the axiom a
  - SpatialRelation is the set of the considered spatial relation
- $R_{DB} = R_G$
- $Rel_{DB} = rel_G$
- $H_{DB} = H_G \cup \{(c,c') | c\in C_G \wedge c' \in C_{DB} \wedge h^{-1}(c')=c\}$

For each new concept added to the DBOntology, a materialized view representing the SQL query correspondent to the axiom defining the concept is added to the database.

As far as the axiom translation into materialized views is concerned, we distinguish different cases, depending on the structure of the selected node. There are three cases: the concept has siblings, the concept has children, and the concept has both siblings and children. We give the intuition of each case with an example, whereas the algorithm is illustrated in the next section.

Case A – Siblings. Consider the PublicBuildingFloodRisk concept that does not have a direct link to the database tables, and its siblings classes, Hospital and School, are directly associated to database tables. The axiom translation results is the creation of a materialized view whose data are the union of the subset of the Hospital and School records located in the flood risk areas.

Reminding that the OWL axiom defining the PublicBuildingFloodRisk concept is *PublicBuildingFloodRisk ≡ PublicBuilding and (intersect some FloodRiskBasin)*, the associated SQL view is:

```
CREATE VIEW C PublicBuildingFloodRisk as
    (SELECT * FROM Hospital  h, FloodRiskBacinhc
     WHERE intersect(h.the_geom, hc.the_geom)
                  UNION[1]
     SELECT * FROM School b, FloodRiskBacinhc
     WHERE intersect(b.the_geom, hc.the_geom))
```

[1] Notice that the use of the UNION operator to include data selected from both tables.

Case B – Children. Let us consider the ontology depicted in Figure 5A and let PedestrianStreet be defined as a Street whose geographical coordinates are located within a PedestrianArea. The correspondent axiom is defined as *PedestrianStreet≡ (Street and (within some PedestrianArea))*. Similarly, suppose RiversideStreet is defined as the streets whose geographical coordinates have a *touch* relation with a river and the correspondent axiom be *RiversideStreet≡ (Street and (touch some River))*. In this example, the concept Street is directly mapped to the database. The axiom translation creates three new materialized viewsthat: (1) select the records of the Street table in a *within* relationship with a PedestrianArea and (2) the records of the Street table located in a *touch* relationship with the River. An additional materialized view, called ComplementStreet, is created to collect the records of the table Street which are not included in the two created views. As a consequence, the union of the records belonging to the three views corresponds to the original Street table, but now we can distinguish between the three cases. The corresponding DBOntologyis depicted in Figure 5B. The SQL statement for the creation of the Complement class is the following

```
CREATE VIEW ComplementStreet as
     SELECT * FROM Street
     WHERE not exist (select * from PedestrianStreet)
     AND not exist (select * from RiversideStreet)
```

Case C – Merge siblings and children. The selected concept has both siblings and children. Both the previous steps are applied in this case.

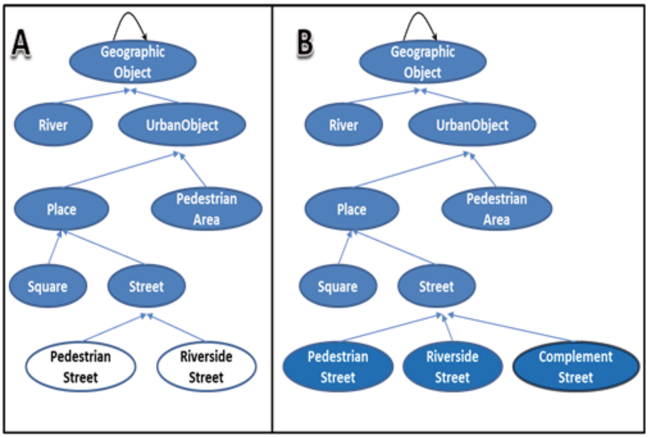


Fig. 5. A) fragment of the urban ontology. PedestrianStreet and Riverside street are define by axioms and not mapped to the database B) Fragment of the corresponding DBOntology. Street concept has been mapped to the database by means of its children PedestrianStreet and Riverside Street. A new concept complement has been added to map the remaining records.

4 The Algorithm of the Enhancing Module

In this section we present the details of the Enhancing Module algorithm. The algorithm is structured in two main phases. The first phase consists in (1) selecting all the domain ontology concepts with a *direct* mapping to database tables, and (2) selecting all the concepts in the domain ontology that are siblings or children of the concepts found at point 1). We call the resulting ontology the DBOntology at step zero (DBO_0). The leaves of this ontology are the concepts mapped directly to database tables. The second phase consists in expanding, as much as possible, the DBO_0 with the input domain ontology concepts defined by axioms, following the rules presented in previous section. At each step *n*, the algorithm selects all the Domain Ontology concepts that are direct children or siblings of the selected DBO_ n-1 leaf, when they are not included in the DBO_n-1. From these concepts the algorithm selects only the ones whose axioms can be solved by an SQL query. This means that we must be sure that all the concepts mentioned in the axioms belong to DBO_n-1. In other words these concepts are directly associated to database tables or already materialized as views. In thelatter case the algorithm creates the corresponding SQL views, inserting the new classes with a *is_a* relationships in the DBO_n-1 thus producing the DBO_n. The process stops when the set of concepts to be materialized is empty, or there is no change from one step to the next one (in this case the remaining concepts will be discarded). The DBO_n constructed by the process is then returned as the final DBOntology.

Let us go through the algorithmic part. Let us recall the formal definition of the ontology given in Section 2.2, and introduce some basic definitions and acronyms used through the algorithm. DBO indicates the DBOntology, DO the Domain Ontology, CA stands for Concept Axiom indicating the set of Domain Ontology concepts defined by axioms and implicitly mapped to database tables. Finally CC indicates ConceptComplement, of the concept to be complemented. A number of predefined functions are used, such as *DefiningConcept DC(c):* is a function which, given a concept defined by an axiom *c*, returns the set of concepts mentioned in the axiom. *hO* indicates the Hierarchy of the ontology O and *hO(c)* returns the set of children of *c*, whereas *hO-1(c)* returns the concept father of *c*. Eventually, *Leaf(O)* is a function that, given an ontology O, returns the set of its leaves. Analogously, *Node(O)* returns the set of node concepts in the ontology O. The function *has_axiom(c)* checks whether a given concept is defined by an axiom. The pseudo-code is illustrated below.

The Enhancing algorithm

```
//Selection of the possible nodes for the enhancing process
  CA = {}
  Leaf = Leaf(DBO) //takes the DBO leaves
  Repeat
    c ∈ Leaf
    c' = $h_{DO}^{-1}$(c) // takes the father of the node
    Leaf\{c}
    forall c'' $\in h_{DO}$(c').c''<>c // the set of c's
```

```
                  //siblings in DO
      if has_axiom(c’’) then
//If the concept is defined by an axiom an does not belong to
the set of DBO nodes and it is not alreadypresent in the set of
leaves, it is added to the CAset//
         if not ((c’’ $\in$ Nodes(DBO)) or
                  (c’’ $\in$ Leaf)) then
            CA $\cup$ {c’’}
  forall c’’ $\in h_{DO}(c)$ // the set c’s children
      if has_axiom(c’’) then
         CA U{c’’}
  until Leaf = {}
//Selection of the concept to be materialized,construction of
the materialized view, update of the DBO, and update of the DB
if needed
  Repeat
      CA_tmp=CA
      Forall c $\in$ CA.
  //takes the concept needed to define the new class
       c' = DC(c)\ $h_{DO}^{-1}$(c)
       if c' $\in$ DBO  then
          CA\{c}
          Create_view(c)
          add_to_DBO(c)
          c'' = $h_{DO}^{-1}$(c)
          if c''$\in$ Leaf(DBO) and not c''$\in$ CC then
             CC U {c''}
  Until CA = CA_tmp
  For all c $\in$ CC Remove_and_Complement(c)
```

The definition of the function *add_to_DBO(c)* follows:

```
  add_to_DBO(c) =
  Leaf = Leaf(DBO)
  Repeat
      c' $\in$ Leaf
      Leaf \ {c’}
  // the father of concept c’ in the HC relation
      c'' $\in$ $h_{DO}^{-1}$(c')
```

```
//c is sibling of c' if $HC_{DO}(c,c'')$ holds
    if $HC_{DO}(c,c'')$ then
}//adds the new concept to the set of DBO concepts
       $C_{DBO} \cup \{c\}$
//adds the father relationship in DBO
        $HC_{DBO} \cup \{(c,c'')\}$
//c is child of c' if $HC_{DO}(c,c')$ holds
    If $HC_{DO}(c,c')$ then
       $C_{DBO} \cup \{c\}$
       $HC_{DBO} \cup \{(c,c')\}$
until $HC_{DO}(c,c'')$ or $HC_{DO}(c,c')$
```

The function *create_view(c)*builds the materialized views as an output of the Enhancing Module. In this code we introduce new functions1) *concept_name(c)* returns the name of the database table the concept *c* is mapped on, 2) *Axiom(c)*returns the axiom that defines the concept *c*, 3) *get_predicate(axiom)*returns the spatial predicate contained in the axiom, 4) *$Leaves_O(c)$* returns all the leaves children of the concept *c* in the ontology O. Furthermore,the function *DBConnection* provides the connection to the spatial DB and executes the query. The *create view(c)*function definition follows.

```
create_view(c) =
//takes the father of c and the other concept in the Enhanced
Ontology needed to define c.
Conc = DC(c)
// takes the set of concepts the father can be rewritten in
ExpFa = {}
$\forall$ c' $\in$ Conc
   if $HC_{DO}(c,c')$ then
      father = c'
//verify if father has to be expanded
      if not isMappedToTable(father) then
         Leaves = $Leaves_{DBO}$(father)
            For all c $\in$ Leaves
        If isMappedToTable(c) then
           ExpFa $\cup$ {c}
        Else
           ExpFa = {father}
        Else
//takes the leafs of the concept c'
         Leaf = $Leaves_{DBO}$(c')
```

```
//production of the query for each Leaf related to the axiom
         Q = {}
         Forall f ∈ExpFa
            Forall c ∈ Leaf
       Q ⋃ {"SELECT * FROM
            " &concept_name(f) & "AS a, "
              &concept_name(c) & "AS b
              WHERE " &
              get_predicate(Axiom(c)) &
             " (a.the_geom, b.the_geom)"}
//creation of the view
query = empty string
q ∈ Q
query = "CREATE VIEW " &concept_name(c) &
        " VIEW as ( " & q
Q\{q}
For all q ∈ Q
   query = query & "UNION " & q
query = query & ")"
//connection to the DB and  view materialization
DBConnection(query)
```

Finally, the definition of the remove_and_complement(c) function follows:

```
Remove_and_complement(c)=
//selection of the children of c in the extended
// enriched ontology
Leaf = Leaves_DBO(c)
query = "CREATE VIEW Complement" &concept_name(c)
         & " as SELECT * FROM " &concept_name(c) &
           " WHERE not exist "
l ∈ Leaf
query = query & "((SELECT * FROM " &
                 concept_name(l) & ")"
Leaf\{l}
Forall l ∈ Leaf
   query = query & " AND not exist
      (SELECT * FROM " &concept_name(l) & " ) " &
query = query & ")"
DBConnection (query)
DBConnection("DROP TABLE " &concept_name(c))
```

5 Related Work

The partnership between ontologies and GIS has seen a growing interest in the last decade [Fonseca02, Mark06, Vidal09, Zaki09]. The role of ontologies in geographical information science can be manifold. A well known topic is Geographical Information Retrieval (GIR), where the ontology supports spatial querying, as witnessed for example by the results of the SPIRIT project [Jones05], where methods for ontology-based spatial query expansion for geographical search engines were studied. For example, in this context, [Cardoso07] proposes a method for the geographical expansion of queries exploiting spatial relationships. Another work [Mata07] proposes the use of an ontology to decide where and what should be searched from different data sources.

Several approaches share with our work the use of ontology for querying geographical databases. However, to the best of our knowledge, no approaches explicitly map ontology axioms to spatial SQL materialized views.

A recent approach [Peachavanish07] proposes a methodology that exploits multiple ontologies for the interpretation of geospatial queries. Compared to our approach, they propose a mediation between ontologies that we are not considering, and, in general, their approach is more conceptual and not based directly on querying database tables. The work in [Torres05] proposes an ontological semantic layer to query a geographical database and in this is similar to our proposal. In particular, their approach allows different community users to access the same geographic database. However, they do not consider specifically the problem of representing the location of a geographical object, neither the translation from ontology axioms to spatial SQL queries which we are handling here.

Similar in the objective, to our approach, the work in [Lüscher08] that aims at enriching the semantics of geodatabase for enhanced user queries. However, compared to the proposed methodology, the authors propose a complementary approach, since they infer semantic information about spatial objects exploiting pattern recognition techniques, that we are not considering here.

The work of [Zhao08] shares with our approach the use of ontologies as a query interface towards spatial data, but the focus there is on data integration and they don't consider the use of a domain ontology to further enhance the geospatial semantics of queries, neither they use a mapping of ontology axioms to spatial SQL queries.

The problem of mapping between ontologies and relational databases is faced by a broad range of literature works. Some example – but the list is not exhaustive, are [Barrasa04, An06, Li05] however, it is out of the scope of the paper to provide a survey on them. Some of these approaches particularly tackle the problem of translating axioms into rules such as [Vasilecas06, Vasilecas07]. However the main differences between these approaches and ours rely on the fact that they do not deal with the spatial domain, and the axioms are not translated in (spatial) SQL materialized views, but in business [Vasilecas06] or ECA rules [Vasilecas07] represented as database triggers. Several other approaches have their roots in the database integration, and they typically map mapping Description Logics formulas to SQL, such as [Levy99] or [Calvanese08].

6 Conclusions and Future Work

In this paper we propose a methodology for mapping ontology axioms to SQL queries on a geodatabase. This approach enhances the usual ontology – to – database mapping, providing support for an automatic translation of ontology axioms to a spatial SQL query. The advantage of this proposal mainly relies on the possibility to support the automatic (or semi-automatic) translation of natural language spatial queries into appropriate database queries stored as materialized views. As a consequence, this methodology increases the number of (ontology) concepts that may be referred in the user queries thus semantically enhancing the query process to a geodatabase.

The advantage of the proposed approach is many-fold: at first it allows to abstract away the domain concepts with respect to the underlying database, thus allowing to reuse the same domain ontology with different database in the same domain. Secondly, we can enhance the user query capabilities exploiting ontology concepts that do not have a classical direct mapping to the database. Finally, the ontology is expressed in a standard Semantic Web technology, such as OWL: it makes easier the use of different domain ontologies and permits different semantical interpretation of the same geographical database.

Some interesting open issue we intend to explore in the future are related to the limitations that we have posed to the present solution. For example, relaxing the limitations on the form of the axioms and generalizing the use of nested relations and object properties.

References

[Acci05] Acciarri, A., Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Palmieri, M., Rosati, R.: QuOnto: Querying ontologies. In: Proc. of the 20th Nat. Conf. on Artificial Intelligence, AAAI 2005 (2005)

[An06] An, Y., Borgida, A., Mylopoulos, J.: Discovering the Semantics of Relational Tables through Mappings. In: Proc. of the 21th Nat. Conf. on Artificial Intelligence, AAAI 2006 (2006)

[Baad03] Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P.: The description logic handbook: Theory, implementation and applications. Cambridge University Press, Cambridge (2003)

[Baglioni07] Baglioni, M., Masserotti, M.V., Renso, C., Spinsanti, L.: Building Geospatial Ontologies from Geographical Databases. In: Fonseca, F., Rodríguez, M.A., Levashkin, S. (eds.) GeoS 2007. LNCS, vol. 4853, pp. 195–209. Springer, Heidelberg (2007)

[Baglioni08] Baglioni, M., Giovannetti, E., Masserotti, M.V., Renso, C., Spinsanti, L.: Ontology-supported Querying of Geographical Databases. Transactions of GIS 12(s1), 31–44 (2008)

[Baglioni09] Baglioni, M., Masserotti, M.V., Renso, C., Soriano, L., Spinsanti, L.: A Tool for Extracting Ontologies from Geographical Databases. In: SEBD 2009, Convegno Italiano di Basi di Dati, Camogli, Genova, Italy (2009)

[Barrasa04] Barrasa, J., Corcho, O., Gómez-Pérez, A.: R2O, An Extensible and Semantically Based Database-to-Ontology Mapping Language. In: Second Workshop on Semantic Web and Databases, SWDB 2004 (2004)

[Bishr08] Bishr, Y.: The Geospatial Semantic Web: Applications. In: Shekhar, S., Xiong, H. (eds.) Encyclopedia of GIS. Springer, Heidelberg (2008)

[Calvanese08] Calvanese, D., De Giacomo, G., Lenzerini, M., Rosati, R.: View-Based Query Answering over Description Logic Ontologies. In: KR 2008, pp. 242–251 (2008)

[Calvanese09] Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Poggi, A., Rodriguez-Muro, M., Rosati, R.: Ontologies and Databases: The DL-Lite Approach. In: Tessaris, S., Franconi, E., Eiter, T., Gutierrez, C., Handschuh, S., Rousset, M.-C., Schmidt, R.A. (eds.) Reasoning Web. LNCS, vol. 5689, pp. 255–356. Springer, Heidelberg (2009)

[Cardoso07] Cardoso, N., Silva, M.J.: Query Expansion through Geographical feature Types. In: GIR 2007, Lisboa, Portugal, November 9 (2007)

[Fonseca02] Fonseca, F.T., Egenhofer, M.J., Agouris, P., Camara, C.: Using Ontologies for Integrated Geographic Information Systems. Transact. GIS 6(3), 231–257 (2002)

[Gru08] Gruber, T.: Ontology. In: Liu, L., Özsu, M.T. (eds.) Encyclopedia of Database Systems. Springer, Heidelberg (2008)

[Guting05] Güting, R., Schneider, M.: Moving Objects Databases. Morgan-Kauffman, San Francisco (2005)

[Levy99] Logic-Based Techniques in Data Integration. In: Levy, A., Minker, J. (eds.) Logic-based Artificial Intelligence (1999)

[Li05] Li, M., Du, X., Wang, S.: Learning Ontology from Relational Database. In: Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, August 18-21 (2005)

[Lüscher 08] Lüscher, P., Weibel, R., Mackaness, W.A.: Where is theTerraced House? On the Use of Ontologies for Recognition of Urban Concepts in Cartographic Databases. In: Ruas, A., Gold, C. (eds.) Headway in Spatial Data Handling, pp. 449–466. Springer, Heidelberg (2008)

[Mark06] Mark, D.M., Egenhofer, M.J., Hirtle, S., Smith, B.: UCGIS Emerging Research Theme: Ontological Foundations for Geographical Information Science (2006)

[Mata07] Mata, F.: Geographic Information Retrieval by Topological, Geographical, and Conceptual Matching. In: Fonseca, F., Rodríguez, M.A., Levashkin, S. (eds.) GeoS 2007. LNCS, vol. 4853, pp. 98–113. Springer, Heidelberg (2007)

[Mata09] Mata, F., Levachkine, S.: iRank: Integral Ranking of Geographical Information by Semantic, Geographic, and Topological Matching. In: IF&GIS 2009, pp. 77–92 (2009)

[OGIS] OpenGIS Simple Feature Access, `http://www.opengeospatial.org/standards/sfa`

[OraSpa] ORACLE, Oracle Spatial, `http://www.oracle.com/database/spatial.html`

[OWL] OWL Web Ontology Language, `http://www.w3.org/TR/owl-features/`

[Peachavanish07] Peachavanish, R., Karimi, H.A.: Ontological Engineering for Interpreting Geospatial Queries. Transactions in GIS 11(1), 115–130 (2007)

[Peuquet02] Peuquet, D.: Representations of Space and Time. The Guilford Press, N.Y. (2002)

[PostGis] PostGres database– PostGIS extension, `http://postgis.refractions.net/`

[Spaccapietra04] Spaccapietra, S., Cullot, N., Parent, C., Vangenot, C.: On Spatial Ontologies. In: GeoInfo (2004)

[Jones05] Fu, G., Jones, C.B., Abdelmoty, A.I.: Ontology-Based Spatial Query Expansion in Information Retrieval. In: Chung, S. (ed.) OTM 2005. LNCS, vol. 3761, pp. 1466–1482. Springer, Heidelberg (2005)

[Torres05] Torres, M., Quintero, R., Moreno, M., Fonseca, F.T.: Ontology-Driven Description of Spatial Data for Their Semantic Processing. In: Rodríguez, M.A., Cruz, I., Levashkin, S., Egenhofer, M.J. (eds.) GeoS 2005. LNCS, vol. 3799, pp. 242–249. Springer, Heidelberg (2005)

[Vasilecas06] Vasilecas, O., Bugaite, D.: Ontology-Based Information Systems Development: The Problem of Automation of Information Processing Rules. In: Yakhno, T., Neuhold, E.J. (eds.) ADVIS 2006. LNCS, vol. 4243, pp. 187–196. Springer, Heidelberg (2006)

[Vasilecas07] Vasilecas, O., Bugaite, D.: An algorithm for the automatic transformation of ontology axioms into a rule model. In: International Conference on Computer Systems and Technologies. ACM International Conference Proceeding Series, vol. 285 (2007)

[Vidal09] Vidal, V.M.P., Sacramento, E.R., de Macêdo, J.A.F., Casanova, M.A.: An Ontology-Based Framework for Geographic Data Integration. In: Heuser, C.A., Pernul, G. (eds.) ER 2009. LNCS, vol. 5833, pp. 337–346. Springer, Heidelberg (2009)

[Zhao08] Zhao, T., Zhang, C., Wei, M., Peng, Z.: Ontology-Based Geospatial Data Query and Integration. In: Cova, T.J., et al. (eds.) GIScience 2008. LNCS, vol. 5266, pp. 370–392. Springer, Heidelberg (2008)

[Zaki09] Zaki, C., Servières, M., Moreau, G.: Combining Conceptual and Ontological Models for Representing Spatio-Temporal Data and Semantic Evolution in GIS. In: Ontologies for urban development: Future development of urban ontologies, Liège, Belgium, pp. 95–104 (2009)

A Supervised Machine Learning Approach for Duplicate Detection over Gazetteer Records*

Bruno Martins

Instituto Superior Técnico, INESC-ID
Av. Professor Cavaco Silva, 2744-016 Porto Salvo, Portugal

Abstract. This paper presents a novel approach for detecting duplicate records in the context of digital gazetteers, using state-of-the-art machine learning techniques. It reports a thorough evaluation of alternative machine learning approaches designed for the task of classifying pairs of gazetteer records as either duplicates or not, built by using support vector machines or alternating decision trees with different combinations of similarity scores for the feature vectors. Experimental results show that using feature vectors that combine multiple similarity scores, derived from place names, semantic relationships, place types and geospatial footprints, leads to an increase in accuracy. The paper also discusses how the proposed duplicate detection approach can scale to large collections, through the usage of filtering or blocking techniques.

1 Introduction

Digital gazetteers are geospatial dictionaries for named and typed places that exist in the surface of the Earth [17]. Their essential utility is to translate between formal and informal systems of place referencing, i.e. between the ad hoc names and qualitative type classifications assigned to places for human consumption, on the one hand, and the quantitative locations (e.g., geospatial coordinates) that support the automated processing of place references, on the other [18].

Digital gazetteers are often built from the consolidation of multiple data sources. Thus, a fundamental challenge with digital gazetteers is record linkage, i.e. detecting exact and near duplicates so that place identity is preserved. Although record linkage techniques have been extensively studied [37], their application to complex data records like in the case of gazetteers (i.e., containing not only textual attributes but also geospatial footprints, name and category multi-sets, hierarchically organized categorical information and semantic relations between the different places), has received less attention in the literature.

By formulating record linkage as a classification problem, where the goal is to classify record pairs as duplicates or non-duplicates, this paper argues that the challenge of gazetteer record linkage can be met by using machine learning, more specifically through supervised classification and feature vectors combining

* This work was partially supported by the Fundação para a Ciência e a Tecnologia (FCT), through project grant PTDC/EIA-EIA/109840/2009 (SInteliGIS).

C. Claramunt, S. Levashkin, and M. Bertolotto (Eds.): GeoS 2011, LNCS 6631, pp. 34–51, 2011.

multiple similarity estimates (e.g. similarity between geospatial footprints, type categories, semantic relations and place names). The paper reports a thorough evaluation on two machine learning approaches designed for the task, from which the following main conclusions can be reached:

- Both support vector machines (SVMs) and alternating decision tree classifiers are adequate to the task, with decision trees performing slightly better.
- Combining different similarity features increases the accuracy, although the similarity between place names alone provides a competitive baseline.
- Similarity scores between place names are the most informative feature for discriminating between duplicate and non-duplicate record pairs.

The rest of the paper is organized as follows: Section 2 presents related work, describing digital gazetteers and approaches for record linkage. Section 3 describes the proposed approach based on supervised classification, detailing the classification techniques and the proposed similarity features. Section 4 presents the experimental validation for the proposed approach. Finally, Section 5 presents the main conclusions and directions for future work.

2 Related Work

This section surveys relevant past research in terms of gazetteer data management and duplicate detection for data consolidation.

2.1 Digital Gazetteers

Gazetteer data exists nowadays in many independent and often dissimilar sources. The place records in modern digital gazetteers comprise different metadata such as the multiple place names, place types, and geospatial footprints corresponding to individual places, to which unique identifiers are also assigned. Since the data involves hierarchies (e.g., hierarchical organizations for the place type categories) and indefinitely repeating groups (e.g., multiple place names or geospatial footprints per individual record), gazetteers are a natural fit for XML technologies. Increasingly, digital gazetteers are accessible through XML Web services, and these services are used in the context of digital libraries and geographic information retrieval (GIR) systems for translating ambiguous place names into unambiguous geospatial coordinates [17,18].

Two of the most relevant past initiatives regarding digital gazetteers are (i) the Open Geographical Consortium's (OGC) gazetteer service interface, and (ii) the gazetteer service developed in the Alexandria Digital Library project.

The Open Geographical Consortium (OGC) defined a gazetteer service interface, i.e. WFS-G, based on a re-factored ISO-19112 content model (i.e., spatial referencing by geographic identifiers) published through a Web Feature Service (WFS[1]). The service metadata, operations, and types of geographic entities are

[1] http://www.opengeospatial.org/standards/wfs

standardized in this specification, which in turns build on the Geography Markup Language (GML[2]) to encode the metadata associated with the places. In particular, WFS-G uses two specific GML feature types, namely SI_Gazetteer and SI_LocationInstance, to encode gazetteer information. Work within the OGC regarding gazetteer services was heavily influenced by the previous work made within the context of the Alexandria Digital Library project.

The Alexandria Digital Library (ADL) was one of the pioneering efforts addressing the development of gazetteer service protocols and data models, mainly to support information retrieval over distributed resources [17]. The ADL gazetteer content standard, i.e. a generic data model for gazetteers, defines the core elements of named places (and their history), their spatial location (in various representations), their relationships to other places (e.g., part of relations), classification (using a referenced typing scheme), and other metadata properties (e.g. source attribution). Regarding place type classification, the ADL project defined an extensive taxonomy of place types, known as the Feature Type Thesaurus (FTT[3]). The current version of the FTT has a total of 1156 place type classes (i.e., terms), of which 210 are preferred terms and 946 are non-preferred terms that are related to the preferred terms. A prototype system describing approximately 5 million U.S. domestic and international places, using the content standard and the FTT, has been deployed by combining the U.S. place names from the Geographic Names Information System (GNIS) and the non-U.S. place names from Geonet Names Server (GNS) gazetteer of the National Geospatial Intelligence Agency (NGA), as well as from other sources.

Within the ADL gazetteer initiative, XML schemas have been proposed for encoding gazetteer data according to the content standard, reusing the geometry schema of the Geographic Markup Language (GML) for encoding geospatial footprints. Web service interfaces for accessing ADL gazetteer data have also been proposed. The datasets used in the experiments reported in this work are all encoded according to the XML schema of the ADL gazetteer protocol, a lightweight version of the ADL gazetteer content standard.

2.2 Duplicate Detection and Data Consolidation

The problem of identifying database records that are syntactically different and yet describe the same physical entity has been referred to as identity uncertainty [28], object identification [24], merge/purge processing [16], record deduplication [32], record linkage [37], or simply duplicate detection [26,27]. Typical methods involve the computation of a similarity score between pairs of records, under the assumption that highly similar records are likely to be duplicates [11]. For every candidate pair of records, a similarity is computed using some distance metric(s) or a probabilistic method. Candidate pairs that have similarity scores higher than a given threshold value can then be linked. The transitive closure of those linked points forms final equivalence classes for the duplicate records.

[2] http://www.opengeospatial.org/standards/gml

[3] http://www.alexandria.ucsb.edu/gazetteer/FeatureTypes/FTT2HTM/

On what concerns similarity measures, past research on duplicate detection has focused on distance metrics computed over strings. Commonly used metrics include the Levenshtein distance [22] (e.g. derived from the minimum number of character deletions, insertions or substitutions required to equate two strings), the Monge-Elkan distance [26] (similar to the Levenshtein distance, but assigning a relatively lower cost to a sequence of insertions or deletions) or the Jaro-Winkler metric [37] (a fast heuristic-method for comparing proper names, which is based on the number and order of the common characters and also accounts with common prefixes). Cohen et al. compared different string similarity metrics, concluding that the Jaro-Winkler metric works almost as well as the Monge-Elkan scheme and is an order of magnitude faster [8]. While the approaches above work well for syntactic matches, duplicate detection should also account with phonetic similarity. The double metaphone algorithm, popularized by the ASpell spelling corrector, compares words on a phonetic basis according to their pronunciation, returning a singular key value for similarly sounding words [21].

Besides string similarity metrics, past research has also addressed the computation of similarity scores between other types of objects. Lin, Resnik and others have suggested semantic similarity measures for categorical information, based on having objects labeled with terms from a hierarchical taxonomy [23,29,33]. On what concerns categorical information based on multi-sets of objects, the Jaccard coefficient measures similarity as the size of the intersection divided by the size of the union of the sample sets. Dice's coefficient also measures similarity between multi-sets of objects, and is defined as two times the size of the intersection divided by the sum of the sizes of the sample sets. Jaccard and Dice's coefficients can also be used as string similarity metrics, by seeing the strings as sets of characters or even as sets of word tokens [8]. Software frameworks such as SimPack [4] implement a wide array of similarity metrics for different types of objects, facilitating the execution of duplicate detection experiments.

Research on duplicate detection has also explored the usage of supervised learning methods, using labelled training data in the form of record pairs that are marked as duplicates or non-duplicates by human editors [5,7,6,32,36,9,35]. Labelled training examples have been explored at two levels, either individually or in combination. These levels are (i) using trainable string metrics, such as learnable edit distance, that adapt textual similarity computations to specific record fields, and (ii) training a classifier to discriminate between pairs of duplicate and non-duplicate records using similarity values for different fields as features. The work reported in this paper falls into the second category, using binary classifiers that learn to combine different features to discriminate duplicate gazetteer records from non-duplicates.

In the machine learning literature, binary classification is the supervised learning task of classifying the members of a given set of objects (in our case, pairs of gazetteer records) into one of two groups, on the basis of whether they have some features or not. Methods proposed in the literature for learning binary classifiers include decision trees [30,12] and support vector machines [25,19]. Decision tree classifiers learn a tree-like model in which the leaves represent classifications and branches represent the conjunctions of features that lead to those classifications. Decision tree classifiers provide high accuracy and transparency (e.g., a human can easily examine the produced rules), although they can only output binary decisions (i.e., the gazetteer record pairs would either be duplicates or non-duplicates). The Alternating Decision Tree algorithm is a generalization of decision tree classifiers introduced by Freund and Mason, which in addition to classifications also gives a measure of confidence (i.e. the classification margin) in the result being the correct classification [12]. Support Vector Machines (SVMs) work by determining an hyper-plane that maximizes the total distance between itself and representative data points (i.e., the support vectors) transformed through a kernel function. SVMs can provide a measure of confidence in the result, i.e. an estimate of the probability that the assigned class is the correct one. This work reports experiments with both alternating decision tree and support vector machine classifiers, through the use of the implementations available in the Weka machine learning framework [38,13].

When detecting duplicate records in the context of data cleaning and data consolidation problems, evaluating all possible pairs of duplicate records is highly inefficient [11]. However, because most of the pairs are clearly dissimilar non-matches, one can design techniques that only select record pairs that are loosely similar (e.g., that share common tokens) as candidates for matching, using blocking [16], canopy clustering [24] or filtering techniques [39,2,1]. These three types of techniques share the fact that they explore computationally inexpensive similarity metrics in order to limit the number of comparisons that require the use of the expensive similarity metrics.

2.3 Duplicate Detection over Gazetteer Records

Previous works have defined the problem of Geospatial Entity Resolution as the process of defining from a collection of database sources referring to locations, a single consolidated collection of true locations [34]. The problem differs from other duplicate detection scenarios mainly due to the presence of a continuous spatial component in geospatial data.

Unlike in the case of place names, which are often associated with problems of ambiguity (e.g., different places may share the same name), geospatial footprints provide an unambiguous form of geo-referencing. Record linkage should therefore be much more trivial in the case of geospatial data and indeed commercial GIS tools use spatial coordinates to join location references, through methods such as the one-sided nearest join in which location references $l_1 \in A$ and $l_2 \in B$ are linked if l_2 is closest to l_1 given all locations in B. However, in practice, data for the spatial domain is often noisy and imprecise. Different organizations often record geospatial footprints using a different scales, accuracies, resolutions and structure [34]. For example, one organization might represent features using only centroid coordinates, while another may use detailed polygonal geometries.

Beeri et al. used geospatial footprints to find matches between datasets, addressing the asymmetry in the definition of the one-sided nearest join [3]. Their results showed that entity resolution is non-trivial even when using only the less ambiguous geospatial information. The use of non-spatial information may improve performance by identifying matching locations which would otherwise be rejected by a spatial-only approach.

Several previous works have proposed to combine both spatial and non-spatial features, although this presents non-trivial problems due to the need for combining semantically distinct similarity metrics. One way to combine different similarity metrics is to put a threshold on one, then using another metric as a secondary filter (i.e. helping in the rejection of similar locations according to the first metric that are not duplicates), and so on [14,15]. Hastings described a gazetteer record linkage approach based on human geocognition, focusing first on geospatial footprints (since overlapping and/or near places can be the same), second on their type categories (conceptually near place types can indicate the same place), and finally on their names (place names with small variations in spelling or with abbreviations, elisions or transliterations, can indicate the same place) [15]. However, the approach above does not capture the matches that are not highly similar according to each of the individual similarity metrics. In this case, one needs a single similarity measure which combines all the individual metrics into a single score. Previous approaches have proposed to use an overall similarity between pairs of features, computed by taking a weighted the average of the similarities between their individual attributes [31]. Weighted averages have the flexibility of giving some attributes more importance than others. However, manually tuning the individual weights can be difficult, and machine learning methods offer a more robust approach.

Zheng et al. proposed a machine learning approach for detecting duplicate records in location datasets, effectively combining features related to name similarity, address similarity and category similarity [40]. Their method was comprised of three steps, namelly candidate selection (i.e., filtering with basis on name similarity and on geospatial distance), feature extraction (e.g., computation of different similarity features) and training/inference based on a decision tree classifier. Experiments with a dataset of 1600 entity pairs consisting of 800 nearly duplicated pairs and 800 non-duplicated ones attested for the advantages of using machine learning for combining different similarity metrics.

Sehgal et al. worked with overall similarity metrics combining footprints (i.e. distance between centroids), place types, and place names (i.e., Levenshtein distance), exploring data-driven approaches using machine learning [34]. This is the most similar work to the research reported in this paper. Sehgal et al. mentioned that, for future work, they would be interested in exploring more semantic information and experimenting with more sophisticated similarity measures. The work reported here goes in this direction, by experimenting with many different similarity metrics computed over a large set of gazetteer record features.

3 Machine Learning for Gazetteer Record Linkage

Classifying pairs of gazetteer records as either duplicates or non-duplicates, according to their similarity, can be seen as a binary, but nonetheless hard, supervised classification problem. Instead of applying a standard record linkage approach, based on string similarity metrics, this paper argues for the use specific similarity features better suited to the context of detecting duplicate gazetteer records. Before detailing the considered features, this section formalizes the considered notion of gazetteer record and presents some examples that illustrate the difficulties in detecting duplicate records.

The general application scenario for the technique reported in this paper is one where we have two gazetteer record datasets A and B developed by independent sources. Each gazetteer record corresponds to some real-world geographic place over the surface of the Earth. A geographic place is defined by (i) one or more place names, by which it is commonly known and communicated, (ii) one or more place types, situating it in an agreed classification scheme that also provides the conceptual basis for it, (iii) zero, one or more geospatial footprints, corresponding to geo-referenced geometries locating it in the Earth's surface and optionally designating its areal configuration, and (iv) zero, one or more temporal footprints, designating the temporal intervals of validity for the place. Each geospatial footprint may be given as a point, line, bounding rectangle, or other polygonal shape that is supported in the Geography Markup Language (GML) standard. Temporal footprints are given as calendar instants or calendar intervals, also according to the GML standard.

Often, a place is given a multiplicity of place names, place types and footprints, by different people and groups, and for different purposes over time. However, within a particular individual gazetteer, a real-world place and its gazetteer record should stand in a one-to-one relationship. Thus, the construction of new gazetteers from multiple data sources requires the handling of duplicate gazetteer records. The objective of gazetteer record linkage is therefore to find pairs of gazetteer records $< r_1, r_2 >$, such that $r_1 \in A$, $r_2 \in B$ and both r_1 and r_2 correspond to the same real world place.

It should be noticed that place names frequently embed place type information (e.g., *Rua Cidade de Roma* or *Luxembourg City*) and some can even be misleading (e.g., *Mississippi*, as a populated place as well as a water body). Different places may share the same or similar centroid coordinates (e.g., *Monaco* the city or the state) and places can change boundaries (e.g., *Berlin* before and after the fall of the wall), place types (e.g., *Rio de Janeiro* was the national capital of *Brazil* before 1960) or place names (e.g., *Congo* was called *Zaire* between 1971 and 1997) over time. The same regions of the globe can also correspond to different places over time (e.g., the former *Union of Soviet Socialist Republics*).

3.1 Gazetteer Similarity Features

The feature vectors used in the proposed record linkage scheme combine information from several different metadata elements available at the gazetteer records. The considered features can be grouped in five classes, namely (i) place name similarity, (ii) geospatial footprint similarity, (iii) place type similarity, (iv) semantic relationships similarity, and (v) temporal footprint similarity.

In terms of the place name similarity features, the idea was to capture the intuition that names that are sufficiently similar support the assessment that a common place is being described. However, the similarity calculations should make allowances for different spellings, abbreviations and elisions, transliterations, etc. Different types of text similarity metrics have complementary strengths and weaknesses. Consequently, it is useful to consider multiple metrics when evaluating potential duplicates. The following similarity features, between place names, were considered in the experiments reported in Section 4:

– The Levenshtein distance between the main place names associated with the pair of gazetteer records being compared.
– The Jaro-Winkler distance between the main place names associated with the pair of gazetteer records being compared.
– The Monge-Elkan distance between the main place names associated with the pair of gazetteer records being compared.
– The Double Metaphone distance between the main place names associated with the pair of gazetteer records being compared. This metric addresses the case in which place names are slightly misspelled.
– The Jaccard coeficient between the sets of alternative names associated with the pair of gazetteer records being compared.
– The Dice coeficient between the sets of alternative names associated with the pair of gazetteer records being compared.
– The minimum and maximum Levenshtein distances between the names given in the sets of alternative names associated with the gazetteer records.
– The minimum and maximum Jaro-Winkler distances between the names given in the sets of alternative names associated with the gazetteer records.

In terms of the geospatial footprint similarity, the idea was to support the intuition that places whose locations are not close cannot be the same. The following features were considered in the experiments:

– The two areas that cover the geospatial footprints associated with the pair of gazetteer records being compared.
– Minimum distance between the geospatial footprints associated with the gazetteer records being compared.
– Distance between the centroid points of the areas that cover the geospatial footprints associated with the records being compared.
– Normalized distance between the centroid points of the areas that cover the geospatial footprints associated with the pair of gazetteer records being compared, given by the formula $similarity = e^{-distance^2}$.
– Area of overlap between the geospatial footprints associated with the pair of gazetteer records being compared.
– Relative area of overlap between the geospatial footprints associated with the gazetteer records, given by Equation 1 originally proposed by Greg Janée[4]:

$$similarity(F_1, F_2) = \frac{area(F_1 \cap F_2)}{area(F_1 \cup F_2)} \tag{1}$$

[4] http://www.alexandria.ucsb.edu/~gjanee/archive/2003/similarity.html

In terms of the place type similarity, the intuition is that places having the same type are more likely to be duplicates. Given the hierarchical nature of the Alexandria FTT, we can also have a sense of proximity between the different place types that are considered (i.e., the place types corresponding to cities and national capitals are conceptually similar and possible descriptors for the same place). The considered set of features is as follows:

- The Jaccard coefficient between the sets of place type classes associated with the pair of gazetteer records being compared.
- The Dice coefficient between the sets of place type classes associated with the pair of gazetteer records being compared.
- The equality of the main place type classes associated with the gazetteer records being compared (i.e., one if they are equal and zero otherwise).
- The semantic similarity metric previously proposed by Lin, computed over the FTT nodes corresponding to the main place type class associated with the gazetteer records being compared.
- The semantic similarity metric previously proposed by Resnik, computed over the FTT nodes corresponding to the main place type class associated with the features being compared.
- The count of the up-steps (to broader terms) necessary to bring the place types to the lowest (narrowest) term that subsumes them both, computed over the FTT nodes corresponding to the main place type class associated with the gazetteer records being compared.

In terms of the similarity measures corresponding to the semantic relations, the intuition is that places related to the same other places are more likely to be duplicates. The considered set of features is as follows:

- The Jaccard coefficient for the set of related features associated with the gazetteer records, for each of the relationship types supported in the ADL gazetteer protocol (i.e., *part of*, *state of*, *county of*, *country of*, etc.).
- The Dice coefficient for the set of related features associated with the gazetteer records, for each of the relationship types supported in the ADL gazetteer protocol (i.e., *part of*, *state of*, *county of*, *country of*, etc.).

Finally, in terms of the temporal similarity, the intuition is that place definitions refering to the same temporal period are more likely to be similar. The following features were considered in our experiments:

- The temporal duration for the interval of overlap between the time periods associated with the pair of gazetteer records being compared.
- The difference between the values for the center dates associated with the pair of gazetteer records being compared.

It should nonetheless be noted that few of the gazetteer records considered in the experiments actually have defined a temporal period for their validity. The impact of these features on the obtained results is therefore neglectable.

In what concerns the implementation of the feature extraction stage, and as perviously stated, all the gazetteer records used in the experiments reported here were encoded in the XML format of the Alexandria Digital Library gazetteer protocol. Small changes were introduced in the XML Schema of the ADL protocol, in order to support the association of gazetteer features to temporal periods (i.e., according to the ADL specifications, temporal information is originally only considered in the full gazetteer content standard).

The loading and parsing of the Alexandria Digital Library gazetteer records was made through the use of the Apache XMLBeans[5] framework. The required geospatial computations were implemented through the use of the Java Topology Suite (JTS), an API of 2D spatial predicates and functions that supports the computation of area aggregates and area intersections [4]. The geospatial footprints are brought to a common datum and projection prior to record linkage, also through the use of JTS. The similarity metrics came from the SimPack package of Java similarity functions [4], except for the double metaphone phonetic distance for which a new implementation was made.

3.2 The Supervised Classification Methods

This paper reports experiments with two different types of classifiers, namely Support Vector Machines (SVMs) and alternating decision tree classifiers. SVMs, based on linear or nonlinear models, represent the state-of-the-art classification technology. They also offer the possibility to assign a value in the interval $[0, 1]$ that estimates the probability of an object (e.g., a pair of gazetteer records) being either classified as positive (e.g., duplicate) or negative (e.g., non-duplicate). These confidence estimates can be viewed as an overall measure of similarity between the gazetteer records comprising the pair. SVMs exhibit remarkable resistance to noise, handle correlated features well, and rely only on most informative training examples, which leads to a larger independence from the relative sizes of the sets of positive and negative examples. Alternating tree classifiers, based on the idea Boosting for combining multiple weak classifiers, can provide more interpretable results (i.e., the decision tree), while at the same time also assigning a confidence value n the interval $[0, 1]$.

Each of these classifiers models the problem with different levels of complexity. For instance, the alternating decision tree classifier tries to define a function that logically partitions the classification space in terms of a tree of decisions made over attributes of the original data, whereas SVMs use a kernel function to flexibly map the original data into a higher-dimensional space where a separating hyper-plane can be defined. A more complex function, such as the SVM approach with a non-linear kernel, may be able to model the training data better, although it can also result in over-fitting the model to the training data.

The feature vectors experimented with both classification algorithms correspond to different combinations of the features described in the previous section. Some experiments were also made with feature selection approaches, for instance measuring the Information Gain statistic associated with the individual features. The Weka machine learning framework provides the implementations for the

[5] `http://xmlbeans.apache.org/`

classification and feature selection algorithms [38,13]. The default parameters in Weka, as associated with the different algorithms, were used in our experiments.

4 Experimental Evaluation

This section describes the datasets, the metrics, and the results for the experimental evaluation of the proposed approach. It also analises filtering techniques for scaling the proposed approach to large sets of gazetteer records.

4.1 The Test Collection of Gazetteer Records

Evaluating the accuracy of a duplicate detection method requires a gold-standard dataset in which all duplicate records have been identified. The experiments reported in this paper used a set of gazetteer records containing both duplicate and non-duplicate examples, having the records encoded according to the XML schema of the ADL gazetteer service. The entire dataset contains 1,257 gazetteer records describing places from all around the globe. A total of 1,927 record pairs have been manually annotated as duplicates. An analysis of the duplicate cases revealed that different situations occur in the dataset, including placenames with different spellings (e.g., *Wien* and *Vienna*), encoded with different geospatial footprints (e.g. either through centroid coordinates, bounding rectangles or complex polygonal geometries) or place types (e.g., *Lisbon* is both a *city* and a *populated place*), and containing more or less detailed information concerning relationships to other place names.

The naive approach of using all possible record pairs described in the dataset (in our case 790,653 record pairs with only 1,927 being duplicates) results in a training set containing much more negative non-duplicate examples than duplicate examples. This skew not only impacts the usefulness of a classification approach (e.g., the accuracy would remain high even if we classified all instances as non-duplicates), but also makes the learning process highly inefficient. Taking inspiration on the approaches proposed by Sehgal et al. [34] and by Bilenko and Mooney [6], the following algorithm was used to select pairs of records to be used in the test collection:

1. All the pairs annotated as duplicates are initially added to the test collection.
2. Assuming that initially we have a set of size n with the record pairs annotated as duplicates, randomly select $n/2$ pairs annotated as non-duplicates and add them to the test collection.
3. The remaining pairs annotated as non-duplicates are sorted according to different similarity metrics, namely the Levenshtein distance on the place names, the centroid distance, the class overlap and the feature type distance.
4. We finally select $n/2$ pairs from the remaining non-duplicates, iterating in a round-robin fashion from the different ordered lists.

Table 1. Characterization of the test collection

Number of records	1,257.00	Records with temporal data	11.00
Number of placenames	1,753.00	Records with only centroid coords.	240.00
Number of unique place names	1,114.00	Total number of pairs	790,653.00
Average names per record	1.39	Total number of duplicates	1,927.00
Number of considered place types	8.00	Considered number of pairs	3,853.00
Average number of types per record	1.11	Considered number of duplicates	1,927.00

The procedure above results in a test collection containing the same number of duplicate and non-duplicate pairs, and it also includes a balance between easy (i.e. highly dissimilar records) and difficult cases (i.e. similar records) for the classifier to decide, the latter being more informative for training than randomly selected record pairs. The total number of record pairs used in the experiments is therefore of 3,853. Table 1 presents a statistical characterization of the test collection and the histograms in Figure 1 show the distribution of the similarity scores for the Levenshtein similarity, the Jaro-Winkler metric, the normalized centroid distance and the relative area of overlap.

4.2 The Evaluation Metrics

A variety of experimental methodologies have been used to evaluate the accuracy of record linkage approaches. Bilenko and Mooney advocate that standard information retrieval evaluation metrics, namely precision-recall curves, provide the most informative evaluation methodology [6]. In this work, precision and recall were computed in terms of classifying the duplicate pairs. Precision is the ratio of the number of items correctly assigned to the class divided by the total number of items assigned to the class. Recall is the ratio of the number of items correctly assigned to a class as compared with the total number of items in the class. Since precision can be increased at the expense of recall, the F-measure (withequal weights) was also computed to combine precision and recall into a single number. Given the binary nature of the classification problem, results were also measured in terms of accuracy and error. Accuracy is the proportion of correct results (both true positives and true negatives) given by the classifier. Error, on the other hand, measures the proportion of instances incorrectly classified, considering false positives plus false negatives.

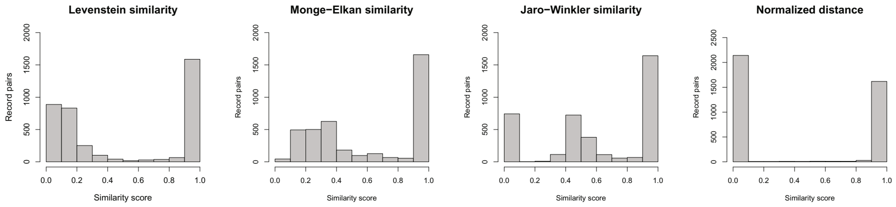


Fig. 1. Histograms with the frequency distribution for different similarity scores

Table 2. Comparison of different classification methods and feature vectors

	Support Vector Machines				
	Precision	Recall	F_1	Accuracy	Error
Placenames	0.993	0.954	0.973	97.379	02.621
Footprints	0.797	0.941	0.863	85.051	14.949
Names+footprints	0.992	0.958	0.975	97.560	02.440
All	0.992	0.962	0.977	97.690	02.310
	Alternating Decision Trees				
	Precision	Recall	F_1	Accuracy	Error
Placenames	0.988	0.971	0.979	97.9496	02.0504
Footprints	0.944	0.950	0.947	94.6535	05.3465
Names+footprints	0.989	0.976	0.983	98.2611	01.7389
All	0.987	0.979	0.983	98.3130	01.6870

4.3 The Obtained Results

Support Vector Machines and alternating decision tree classifiers were trained using different combinations of the proposed features. The considered feature combinations are as follows:

1. Use only the place name similarity features.
2. Use only the geospatial footprint similarity features.
3. Use the place name and the geospatial footprint similarity features.
4. Use the entire set of proposed similarity features.

In each of the above cases, a 10-fold cross validation was performed. Table 2 overviews the results obtained for each of the different combinations. The results show that the alternating decision tree classifiers consistently outperform the classifiers based on support vector machines. A decision tree classifier that used all the proposed similarity features achieved the top performance (i.e., an accuracy of 98.313), although the usage of place name similarity alone provides a very competitive baseline (i.e., an accuracy of 97.9496). The top performing classifier corresponds to a decision tree with 41 nodes and 21 leafs, using 9 different features. The root of the decision tree tests if the overlap coefficient between the sets of place names is less or more than zero.

Using the geospatial footprint similarity metrics alone results in an accuracy of 94.6535 and the combination of the name plus the footprint similarities results in an accuracy of 98.2611. The remaining types of similarity features (e.g., place type similarity) seem to have a limited impact on the final results.

Although using place name similarity alone results in an accuracy that is close to the best reported combination of features, some caveats should be attached to this conclusion. The collection of gazetteer records that was used in the experiments only contains relatively high-level places (e.g. cities). Since place name ambiguity is likely to itself manifest more over thin-grained places (e.g., small villages or street names), the performance of place name similarity should drop in the case of gazetteer records containing thin-grained information. Using the

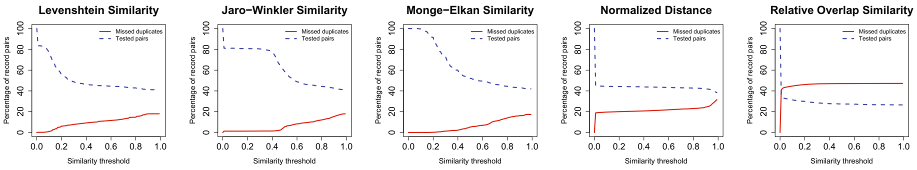


Fig. 2. Effect of pre-filtering pairs with basis on simple similarity scores

information gain statistic, a specific test examined which of the proposed features are the most informative. For each of the four scenarios considered in Table 2, Table 3 lists the top five most informative features. The results show that indeed place name similarity provides the most informative features in discriminating between duplicate and non-duplicate gazetteer records.

Feature selection has a long history within the field of machine learning. Through appropriate feature selection, one can often build classifiers that are both more efficient (i.e., using less features) and more accurate. A specific experiment address the use of a greedy forward feature selection method. Greedy forward selection begins with a set of pre-selected features S which is typically initialized as empty. For each feature f_i not yet in S, a model is trained and evaluated using the feature set $S \cup f_i$. The feature that provides the largest performance gain is added to S, and the process is repeated until no single feature improves performance. An alternating decision tree classifier was build through this procedure. The results obtained with this classifier correspond to an accuracy of 97.4306, considering a total of eight features. Of these eight features, two are related to the similarity between spatial footprints and one concerns with place types. The remaining features correspond to place name similarity metrics.

A last experiment examined how the proposed categorization approach can scale to large collections, through the usage of pre-filtering techniques. The idea is that, instead of comparing all pairs, we can pre-filter the pairs to be compared according to individual similarity scores produced by highly informative and computationally inexpensive approaches. The charts in Figure 2 show, for five different pre-filtering techniques, how many duplicate records would be missed, if we only compared record pairs having a similarity score above a given threshold.

The results show that the both the Levenshtein and Jaro-Winkler metrics, computed over the primary place names associated with the gazetteer records, provide good pre-filtering approaches. For instance, if we consider only the records pairs where the Levenshtein similarity is above 0.9, automatically classifying all other

Table 3. The top five most informative features

1 - Names	2 - Footprints	3 - Names+Footprints	4 - All
Jaccard Names	Centroid Distance	Jaccard Names	Jaccard Names
Overlap Names	Distance	Overlap Names	Overlap Names
Levenshtein Primary	Normalized Distance	Levenshtein Primary	Levenshtein Primary
JaroWinkler Primary	Relative Overlap	JaroWinkler Primary	JaroWinkler Primary
MongeElkan Primary	Overlap	Centroid Distance	Centroid Distance

pairs as non-duplicates, we would only be missing around 15.4 percent of the full set of duplicate pairs. At the same time, the number of pairs to compare through the computationally more expensive procedure involving the use of the full set of similarity features would be reduced to around 41.8 percent of the entire set of record pairs. Computing thresholds on the normalized distance and on the relative overlap is also relatively inexpensive, particularly if one considers appropriate indexing mechanisms. However, they provide much worse pre-filtering heuristics.

5 Conclusions and Future Work

This paper presented a novel approach based on supervised learning for finding duplicate gazetteer records. It reported a thorough evaluation of two different classification approaches (i.e., Support Vector Machines and alternating decision tree classifiers) using feature vectors that combine different aspects of similarity between pairs of gazetteer records. These aspects are (i) place name similarity, (ii) geospatial footprint similarity, (iii) place type similarity, (iv) semantic relationship similarity, and (v) temporal footprint similarity. Both SVMs and alternating decision tree classifiers are adequate to the task, with alternating decision trees performing slightly better. The usage of all the different types of similarity features leads to an increase in accuracy, although the similarity between place names is the most informative feature.

Despite the promising results, there are also many challenges for future work. Previous studies have acknowledged that duplicate detection can be further complicated by the fact that the different data sources may use different vocabularies for describing the location types, motivating the use of semantic mappings for cross-walking between the classification schemes [34]. In this work, it was assumed that the gazetteer records were all using the classification scheme of the ADL feature type thesaurus and we limited the experiments to using similarity metrics between nodes in this thesaurus. An interesting direction for future work would be to explore the use of a similarity metric that accounted with the semantic differences between the feature types. It should nonetheless be noted that feature type similarity scores were not among the most informative features, and it seems reasonable to assume that it is possible to accurately identify duplicates even without a common classification scheme.

Previous works have also noted that the standard textual similarity metrics are not well suited to place names because, in everyday usage, their stylistic variability is too great [14]. Also, at the token level, certain words can be very informative when comparing two strings for equivalence, while others are ignorable. For example, ignoring the substring *street* may be acceptable when comparing address names, but not when comparing names of people (e.g. James Street). Advanced string similarity metrics specific for geographic names have been proposed in the past [14,10] and it would be interesting to integrate these metrics, as additional features, in the proposed machine learning framework.

When detecting duplicates, besides the information that is directly available in the pairs of gazetteer records being compared, information available on related gazetteer records may also prove useful useful. For instance, in two dispar sources, duplicate places may be described through records with completely different names. However, the majority of the gazetteer records associated with the two will share many names in common. Similarly to Samal et al. [31], future work could address the extension of the pairwise similarity between features in order to consider contextual information given by related features with a high semantic similarity (i.e., associated through part-of relationships) or geographic proximity (i.e., through centroid distance). It should nonetheless be noted that even the relatively simple metrics used in the experiments reported here already provide very accurate results.

Finally, previous works have also addressed semi-automated gazetteer record linkage, through a user-interface specifically designed for the task [20]. Currently ongoing work also goes in this direction, by studying user interfaces for the management of gazetteer information in which human editors are asked to access the results of automated approaches for gazetteer record linkage, and latter are asked on how to perform record fusion. Fusion is particularly hard, since it has to deal with problems of inconsistency, redundancy, ambiguity, and conflictive information in the collection. The overall objective is to have redundancies eliminated, accurate data retained and data conflicts reconciled, in building useful gazetteer datasets resulting from the fusion of multiple heterogeneous sources.

References

1. Arasu, A., Ganti, V., Kaushik, R.: Efficient exact set-similarity joins. In: Proceedings of the 32nd International Conference on Very Large Data Bases (2006)
2. Bayardo, R.J., Ma, Y., Srikant, R.: Scaling up all pairs similarity search. In: Proceeding of the 16th International Conference on World Wide Web (2007)
3. Beeri, C., Kanza, Y., Safra, E., Sagiv, Y.: Object fusion in geographic information systems. In: Proceedings of the 30th International Conference on Very Large Data Bases (2004)
4. Bernstein, A., Kaufmann, E., Kiefer, C., Bürki, C.: Simpack: A generic java library for similiarity measures in ontologies (2005) (working paper)
5. Bilenko, M., Kamath, B., Mooney, R.J.: Adaptive blocking: Learning to scale up record linkage and clustering. In: Proceedings of the 6th IEEE International Conference on Data Mining (2006)
6. Bilenko, M., Mooney, R.J.: On evaluation and training-set construction for duplicate detection. In: Proceedings of the KDD 2003 Workshop on Data Cleaning, Record Linkage, and Object Consolidation (2003)
7. Bilenko, M., Mooney, R.J.: Adaptive duplicate detection using learnable string similarity measures. In: Proceedings of the 9th ACM Conference on Knowledge Discovery and Data Mining (2006)
8. Cohen, W., Ravikumar, P., Fienberg, S.: A comparison of string distance metrics for name-matching tasks. In: Proceedings of 9th ACM Conference on Knowledge Discovery and Data Mining (2003)
9. Cohen, W.W., Richman, J.: Learning to match and cluster large high-dimensional datasets for data integration. In: Proceedings of 8th ACM Conference on Knowledge Discovery and Data Mining (2002)

10. Davis, C., Salles, E.: Approximate string matching for geographic names and personal names. In: Proceedings of the 9th Brazilian Symposium on GeoInformatics (2007)
11. Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S.: Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering 19(1) (2007)
12. Freund, Y., Mason, L.: The alternating decision tree learning algorithm. In: Proceedings of the 16th International Conference on Machine Learning (1999)
13. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explorations Newsletter 11 (2009)
14. Hastings, J., Hill, L.L.: Treatment of duplicates in the alexandria digital library gazetteer. In: Proceedings of the 2002 GeoScience Conference (2002)
15. Hastings, J.T.: Automated conflation of digital gazetteer data. International Journal Geographic Information Science 22(10) (2008)
16. Hernandez, M.A., Stolfo, S.J.: The merge/purge problem for large databases. In: Proceedings of the 1995 ACM Conference on Management of Data (1995)
17. Hill, L.L.: Core elements of digital gazetteers: Placenames, categories, and footprints. In: Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries (2000)
18. Hill, L.L.: Georeferencing: The Geographic Associations of Information. The MIT Press, Cambridge (2006)
19. Joachims, T.: Making large-scale SVM learning practical. In: Scholkopf, B., Burges, C.J.C., Smola, A.J. (eds.) Advances in Kernel Methods - Support Vector Learning. The MIT Press, Cambridge (1999)
20. Kang, H., Sehgal, V., Getoor, L.: Geoddupe: A novel interface for interactive entity resolution in geospatial data. In: Proceeding of the 11th IEEE International Conference on Information Visualisation (2007)
21. Lawrence, P.: The double metaphone search algorithm. C/C++ Users Journal 18(6) (2000)
22. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10 (1966)
23. Lin, D.: An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference on Machine Learning (1998)
24. McCallum, A.K., Nigam, K., Ungar, L.: Efficient clustering of high-dimensional datasets with application to reference matching. In: Proceedings of 6th ACM Conference on Knowledge Discovery and Data Mining (2000)
25. Moguerza, J.M., Muñoz, A.: Support vector machines with applications. Statistical Science 21(3) (2006)
26. Monge, A.E., Elkan, C.: The field matching problem: Algorithms and applications. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (1996)
27. Naumann, F., Herschel, M., Ozsu, M.T.: An Introduction to Duplicate Detection. Morgan & Claypool Publishers (2010)
28. Pasula, H., Marthi, B., Milch, B., Russell, S., Shpitser, I.: Identity uncertainty and citation matching. In: Proceedings of the 7th Annual Conference on Neural Information Processing Systems (2003)
29. Resnik, P.: Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research 11 (1999)
30. Safavian, S.R., Landgrebe, D.: A survey of decision tree classifier methodology. IEEE Transactions on Systems, Man and Cybernetics 21(3) (1991)

31. Samal, A., Seth, S., Cueto, K.: A feature-based approach to conflation of geospatial sources. International Journal of Geographical Information Science 18 (2004)
32. Sarawagi, S., Bhamidipaty, A.: Interactive deduplication using active learning. In: Proceedings of 8th ACM Conference on Knowledge Discovery and Data Mining (2002)
33. Schwarz, P., Deng, Y., Rice, J.E.: Finding similar objects using a taxonomy: A pragmatic approach. In: Proceedings of the 5th International Conference on Ontologies, Databases and Applications of Semantics (2006)
34. Sehgal, V., Getoor, L., Viechnicki, P.D.: Entity resolution in geospatial data integration. In: Proceedings of the 14th International Symposium on Advances on Geographical Information Systems (2006)
35. Tejada, S., Knoblock, C.A., Minton, S.: Learning domain-independent string transformation weights for high accuracy object identification. In: Proceedings of 8th ACM Conference on Knowledge Discovery and Data Mining (2002)
36. Winkler, W.E.: Methods for record linkage and bayesian networks. Technical report, Statistical Research Division, U.S. Census Bureau (2002)
37. Winkler, W.E.: Overview of record linkage and current research directions. Technical report, Statistical Research Division, U.S. Census Bureau (2006)
38. Witten, I.H., Frank, R.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco (2000)
39. Xiao, C., Wang, W., Lin, X., Yu, J.X.: Efficient similarity joins for near duplicate detection. In: Proceeding of the 17th International Conference on World Wide Web (2008)
40. Zheng, Y., Fen, X., Xie, X., Peng, S., Fu, J.: Detecting nearly duplicated records in location datasets. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems (2010)

An Approach to the Management of Multiple Aligned Multilingual Ontologies for a Geospatial Earth Observation System

Kristin Stock and Claudia Cialone

Centre for Geospatial Science, University of Nottingham, UK
{Kristin.Stock,Claudia.Cialone}@nottingham.ac.uk

Abstract. Ontologies are widely used, within and outside the geospatial context to support semantic search that is capable of returning suitable resources. Some large, heterogeneous earth observation systems that are currently being developed in a multi-thematic environment require the support of multiple ontologies. Furthermore, some of the systems under current development operate in a multi-lingual environment, and it is desirable that multiple languages be supported by the systems themselves.

This paper proposes a solution to this set of requirements using an architecture containing multiple and multilingual ontologies. Such ontologies are required to be related and the architecture described in this work, which adopts a spatial data infrastructure based on open geospatial standards, employs an algorithm for semantic search across the multiple multilingual ontologies aligned using the W3C Simple Knowledge Organization System (SKOS). It also provides an approach that is extendable by the addition of further ontologies if they are required for particular thematic purposes. A number of issues arose during phases of implementation, but the broad approach proved effective for supporting a large, heterogeneous, multilingual earth observation system.

Keywords: multilingual ontologies, semantic alignment, natural language query, geospatial systems, interoperability.

1 Introduction

Description and discovery of resources (for example, web services and data sets) in geospatial earth observation systems are often supported by controlled vocabularies. These sets of concepts or terms are useful in that they allow users to describe their resources using a known set of terms that can be used for querying. If a completely free keyword approach is used instead, resource discovery is often less effective because users may not query using the terms that resource providers have used to describe their resources.

Controlled vocabularies are also sometimes augmented with the definition of semantic relations between concepts, in which case they may be referred to as thesauri. If they further include logical constraints, are defined by groups of individuals sharing a conceptualization and/or are defined in a particular formal language, they may also be referred to as ontologies [1].

C. Claramunt, S. Levashkin, and M. Bertolotto (Eds.): GeoS 2011, LNCS 6631, pp. 52–69, 2011.

Ontologies are often defined by an information community for a particular purpose, are sometimes thematically based and the concepts and their definitions are often limited to one or a small group of human languages. However, large geospatial information systems often cover a large range of information communities covering multiple languages and themes. It is difficult to define a single ontology across multiple-themes, because different groups use different conceptualizations and terminologies, different levels of detail in different topic areas, and speak different languages. Thus it is often the case that more than one ontology is required. However, such multiple ontologies must be used together and related to one another in order for them to effectively support heterogeneous information systems appropriately.

This paper describes an approach to the management of multiple ontologies that allows existing ontologies to be retained in their original format, but that permits theoretically any number of ontologies to be added to a system to cover particular thematic areas, languages or conceptualizations. This approach uses ontology alignment to define semantic relations between concepts from different ontologies. The paper presents the approach and discusses the issues involved in such an undertaking. This work extends previous ontology alignment activities by other researchers in that it considers the pragmatic issues involved in large scale ontology alignment with thematic experts and describes how ontology alignment fits within a broader geospatial earth observation system.

The outline of the paper is as follows: Section 2 describes a case study around which the work is based; Section 3 describes related work; Section 4 presents the approach, including how the mappings are defined and the architecture of the system that implements it; Section 5 discusses issues that arose, evaluates the work and discusses further research.

This paper will give the reader an appreciation of the issues involved in the management of different semantic structures for the support of large geospatial information systems and infrastructures involving multiple languages and multiple themes, including the necessary trade-offs between the representation of multiple conceptualizations and the complexities that such representations create.

2 The Case Study: EuroGEOSS

The work described in this paper was undertaken in the context of the European Commission-funded EuroGEOSS project.[1] EuroGEOSS aims to improve the scientific understanding of the complex mechanisms affecting our planet by establishing interoperable arrangements between environmental information systems.

EuroGEOSS tries to achieve interoperability in a number of ways, including: a) querying based on web 2.0 principles[2]; b) workflow approaches to integrate different scientific models;c) semantic discovery of drought resources using visual and alignments of multilingual ontologies[3]. The latter implementation particularly makes use of a semantic discovery augmentation component where the multilingual aspect is treated by using the ontological multilingual options. The work described in this paper

[1] EUROGEOSS website: `http://www.eurogeoss.eu/default.aspx`

[2] University of Zaragoza is developing this.

[3] More info: GeoS2011 ‘Inter-disciplinary interoperability for global sustainability research’.

could be seen as a semantic and linguistic extension to all this and another aspect of interoperability, based on the use of recursive retrieval of resources using formal aligned ontologies and multilingual analysis as a foundation for later work on natural language querying of spatial relations that other approaches lack.

EuroGEOSS makes use of the specifications of already existing systems such as GEOSS (the Global Earth Observation System of Systems)[4] and INSPIRE (the Infrastructure for Spatial Information in the European Community).[5]

The EuroGEOSS project infrastructure focuses on the retrieval, discovery and harmonization of a large amount of environmental data in three thematic areas: forestry, drought and biodiversity. Data is available at local, regional and global levels covering these strategic areas. The challenge is to render it semantically and technically interoperable in a simple way. The approach taken by EuroGEOSS involves the selection of supportive ontologies for the development of more attainable semantic data interoperability, within Spatial Data infrastructures (SDIs) and for the Semantic Web. The particular challenges of EuroGEOSS that are relevant to the work described in this paper were that: a) three different themes were involved, each complex and with its own existing information communities and requiring different levels of detail; b) the system covered Europe, and so involved many natural languages c) there was no existing single ontology that was appropriate for the task: some were too general, too specific, or not suited to high level browsing, and some not multilingual.

The case study indicated the need for an approach to the management of multiple and multilingual ontologies in an extendable and practical way. The work described supports EuroGEOSS in that it enhances a semantic and multilingual system that renders thematic interoperability possible in a number of ways.

3 Related Work

The approach adopted in this paper reflects the needs of large, multi-theme, multilingual geospatial systems like EuroGEOSS. EuroGEOSS follows and extends other already implemented or ongoing projects.

The Global Earth Observation System of Systems (GEOSS)[6] is a multi-disciplinary and international approach addressing big global environmental issues. It makes use of a new spatial data infrastructure that connects data- providers with users to endorse societal benefit. GEOSS is supported by its own vocabulary that addresses crucial areas of societal benefit (for example, energy, agriculture, biodiversity), but this vocabulary is not multilingual.

The Generic European Sustainable Information Space for the Environment (GENESIS) project[7] originated in 2008 at the EU Joint Research Center. It is planned to provide Europe with a web-based infrastructure to monitor air and water and their impacts on human health. GENESIS has created a semantic repository to store and

[4] GEOSS: `http://www.earthobservations.org/`

[5] INSPIRE: `http://inspire.jrc.ec.europa.eu/`

[6] GEOSS: `http://www.epa.gov/geoss/index.htm`

[7] GENESIS: `http://genesis-fp7.eu/`

link knowledge schemes and a tool (SKOS Matcher)[8] that generates machine-readable/processable files from users' supervised mappings. However, this project does not address the multilingual issue from a user perspective.

The GIS4EU project[9] involves 23 European countries and their institutions and is mainly focussed on data sharing and integration providing cartography datasets for Europe in the areas of administrative units, hydrography, transportation networks and terrestrial elevation. Geospatial ontologies have been generated to enhance interoperability and semantic reference among heterogeneous geospatial information systems.

A further project, FinnONTO (2003–2012) project, describes an approach for the management of multiple ontologies that does not try to solve interoperability problems but tries to avoid them by means of a synergetic collaboration in developing open source vocabularies/ontologies [2].

None of these projects addresses the issue of alignment/creation of multiple ontologies in a multilingual environment.

3.1 Ontology Merging vs. Ontology Alignment

When a multiple ontology-based infrastructure is considered, the concepts included in each ontology are required to be related to each other. Currently, there are two main methods to do this. These are ontology merging, which has been defined as the recreation of a single ontology as an integrated version of the original sources covering similar or overlapping domains; and ontology alignment (in this paper also described as mapping) defined as the creation of consistent and coherent relations between two autonomous original ontologies covering domains that are complementary to each other [3][4].

Ontology alignment in general has the main advantage of maintaining the basic autonomy of the sources. However, one of the drawbacks of alignment is the lack of synchronized and centralized control [5]. Moreover another problem lies in the nature of the ontologies themselves distinguished, depending on their semantic granularity, as shallow or deep ontologies [6][7]. This may affect badly the outcome of the queries. Therefore, to avoid mismatches, a manual contribution and a considerable amount of time are required to be exerted by the users to ensure that concepts are aligned correctly, prior to making any query. On the other hand, while merging makes the process of querying easier, it requires a single conceptualization, and thus valuable concepts with totally different meanings could be irremediably lost [8].

The current approach opted for the alignment technique responding to the very complex nature of the geospatial system it operates in. This in fact requires more flexibility and decentralization in its semantic structure.

3.2 Approaches to Automatic Ontology Matching

In the case of both ontology alignment and ontology merging, automatic methods for ontology matching may be employed. These approaches involve determining the

[8] JRC SKOSMatcher: `https://semanticlab.jrc.ec.europa.eu/SKOS Matcher/`. Note that access to the semantic repository requires private identification from project partners.

[9] GIS4EU: `http://www.gis4eu.eu/`

semantic distance between concepts [9], and include totally automatic methods [10]; graphical approaches [7]; semi-automated methods [11]; and visual tools [12][13].

Although they may make the matching process easier, these approaches present limitations related to the multiplicity of purposes of each matching. This means that a global algorithm efficient for multiple scenarios still does not exist, given the multiplicity of attempts in different fields of application. For example some academic communities may build a 'Water ontology' for chemical purposes and refer to it as 'substance' thus describing its chemical components (molecular composition, PH) and so on. Other geographical communities might want to build a 'water ontology' to organize their knowledge around the concept of 'water' as a geographic feature, water on Earth, including descriptions of hydrographic basins (such as lakes, rivers), effects on flora and fauna etc. Inevitably the two ontologies will present a considerably high number of similar terms whose description varies according to their context.

Therefore, an automatic matching is discouraged as a careful understanding of the perspective adopted by the two communities is required. Cruz et al. also address the problem, especially in the geospatial context, of the divergence in the topological organization of knowledge [14][15]. However, most of these approaches do require human (expert) verification or a preliminary manual matching in any case, constituting an extra exercise for the users, since machines lack critical and common sense.

The current work, instead, does not address the issue of automatic matching, for it evaluates much more a manual alignment of already existing or new small ontologies, given the number of contextual meanings that each ontological scheme can have. However, an automated or semi-automated matching approach, when required with a more complex set of ontologies, could easily be combined with the discussed approach , proposed algorithm and architecture described in this paper.

3.3 Ontology-Based Querying with Multiple Ontologies

Thus far, investigation into ontology based querying has pointed to the possibility for users to concentrate on ontologies as simple keyword clusters to assist them in extracting resources during their discovery. Current approaches have been adopted to solve a number of hindrances that have been encountered [16][17][18]. These approaches put forward crucial issues (multilingualism, localization and semantic heterogeneity, cognitive diversity) but very few talk about concrete mapping procedures 'across' different ontologies in a decentralized spatial data infrastructure. One of the challenges behind the adoption of a multiple-ontology approach is to expand the users' queries across the different ontologies in an effective and efficient way. Lacasta et al.[19] address this issue in an Open Geospatial Consortium (OGC)[10], with a Web Ontology Service (WOS) architecture. They see the biggest challenge in overcoming conceptual and multilingual barriers between the semantics adopted by the users' queries and the one at the disposal of the different service providers. They propose starting with the semantic and linguistic input of the users matching it with as many synonyms as possible from related multilingual ontologies such as GEMET, AGROVOC and EUROVOC.

[10] OGC: `http://www.opengeospatial.org/`

Another approach applied to the geographic field is that proposed by Vidal et al. [20]. This is aimed at increasing the percentage of discovery responses over a heterogeneous system and is based on refining users' queries.

Other work has addressed the problem of extending the queries' semantics over a number of different geospatial ontologies, including users' domain ontologies, a geographic domain ontology (for a wider generalization) and a top-level ontology (the latest general conceptualization) [21].

In the above-cited approaches, and generally not many recent ones in the geospatial literature, the issue of querying multiple ontologies in a multilingual and multi-thematic environment has not been fully addressed. Some of them in fact, tend to give more space to monolingual environment and those who address multilingualism remain within fixed domain limits of application. An attempt to apply a novel approach to span between the two is described in this work.

4 The Approach

This paper proposes an approach to the management of ontology alignment that can be used to support multi-lingual queries across a range of environmental themes. Moreover, it is extensible and thus allows users to add their own multilingual ontologies to cover specialized areas with more detailed concepts not included in the core ontologies. This Section describes the approach in detail, beginning with the core ontologies and the alignment that was performed between them, followed by the ways in which users can add new ontologies in their areas of interest and a presentation of the architecture and the use of the approach to support ontology-based querying.

4.1 The Core Ontologies

Two core ontologies were selected to provide a foundation for knowledge representation and ontology querying in the EuroGEOSS project. The number of core ontologies was limited to two because of resource constraints, particularly among thematic experts. The criteria for selection of the core ontologies were that they should: a) conform to and comply with standards; b) support multiple-languages; c) be available in SKOS, RDF[11] or OWL[12] format although SKOS was preferred given its simplicity and its power to interlink data across the Web; d) provide a broad set of high-level environmental concepts e) provide concepts covering the focus areas of the EuroGEOSS project: drought, forestry and biodiversity.

On the basis of these criteria, two existing, autonomous domain ontologies (providing the basic environmental vocabulary suitable to cover and to annotate resources from the thematic areas of concern) were selected: the GEOSS Societal Benefit Area (SBA)[13] categories and subcategories and GEMET.[14]

[11] RDF: http://www.w3.org/RDF/

[12] OWL http://www.w3.org/2004/OWL/

[13] SBAs in GEO portal website:
http://www.geoportal.org/web/guest/geo_home

[14] GEMET: http://www.eionet.europa.eu/gemet

The SBA categories and subcategories is a minimal, top-level, vocabulary of 9 categories and 58 subcategories addressing environmental issues of global interest. The SBAs were created under the GEOSS project.

Originally, only an English version of the Societal Benefit Areas was provided. However, in order to accommodate the specific multilingual query opportunities needed for the EuroGEOSS project, we have developed versions of the SBAs terms in Italian, Spanish, French, and Slovenian, and these translations are the subject of a future publication by the authors.

GEMET is a multilingual thesaurus developed under of the egis of the European Environment Agency (EEA). It can be conceived of as a middle-level ontology with over 6,500 descriptors related to the environment, provided in 23 official European languages. GEMET is divided into three super categories, 30 subcategories and 34 Spatial Data Themes. GEMET presents a structure that includes both vertical (hierarchical) and horizontal (associative) relations between its concepts.

The SBA categories and subcategories represent another level of generalization than GEMET and for this reason they can provide a complementary terminological umbrella with the GEMET terminology.

4.2 Multilingualism

Exposing more than one language was one of the prerequisites in selecting the ontologies for our approach. For this reason, we had to translate the SBAs. The translation of the vocabulary was accomplished by the University of Nottingham and inputs by international members of the project for French, Italian and Spanish versions and the University of Ljubljana provided a Slovenian version.[15]

The translation of the SBAs proved difficult to undertake since not only did it involve a linguistic knowledge of the technical terms but also a cultural knowledge of how these terms are currently used. The biggest challenge was the consideration of the hierarchy in the vocabulary. English for example might have terms whose translation in another language (for example, Spanish) corresponds to a subcategory of that same English term for which no correspondence in Spanish might exist. For example, the English term ‘pollution events’, does not have an exact match in Spanish. Therefore, one is forced to adopt translations such as ‘desastre ecológico’, or ‘marea negra’ which are respectively a supercategory and a subcategory of the term ‘pollution events’, and not an exact translation.

Furthermore ontologies by default assign to any concept a unique URI. The system takes advantage of this and in its internal engine it uses the URIs of the ontology concepts and not the terms (or labels) themselves to retrieve resources. This is possible because of the functionalities of metadata editors used to tag resources through keywords both as URIs and terms. The reason behind this choice is due to the fact that every term in an ontology has a URI as a universal identifier to which a number of different languages refer. Therefore, given a term selected by the user in any language supported by the ontologies (multilingual to facilitate this aim), the system can retrieve any resource in any language tagged with that URI behind that term selected.

[15] Societal Benefit Areas translations:
`http://en.wikipedia.org/wiki/Societal_Benefit_Areas`

4.3 Alignment of the Core Ontologies

The core ontologies were aligned manually using SKOS[16] as part of the project so that they could be used in combination for querying and resource annotation. In some cases, GEMET may be most appropriate for a particular purpose, while in other cases, the SBA categories and subcategories may be better. To give some examples, the term 'water (geography)' in GEMET, which represents a general domain keyword might be used, by experts, for addressing general resources related to hydrographic features (e.g. seas, lakes, rivers etc). On the other hand, the term 'water cycle research' in the GEOSS SBAs could be used to address more specific resources related to theoretical research conducted on water.

SKOS provides a set of semantic relations used to define relationships within an ontology, and a set of mapping properties used to define relationships across ontologies. The mapping properties are: a) skos:broadMatch and skos:narrowMatch, indicating a hierarchical relation; b) skos:relatedMatch, used to state an associative mapping between two sister concepts; c) skos:closeMatch, which links two concepts that are sufficiently similar to be used interchangeably in some information retrieval applications but not always; d) skos:exactMatch, which links concepts that can be used interchangeably across a wide range of information retrieval applications.

One of the issues from a semantic point of view is the inability to directly express overlapping relations in SKOS. In two concepts that have an overlapping relation, each concept shares some of the members of the other concept, but not all. In some cases, these concepts may be used interchangeably but not in all cases [22] given the difference in semantic features that one term possesses while the other does not [23]. In SKOS, the skos:closeMatch relation may be used to reflect overlapping relations, but it may also describe other types of relations. This semantic ambiguity can limit the deductions that can be drawn regarding relations between concepts in different ontologies.

After a first graphical alignment refined by thematic experts, the manual SKOS mappings were defined with the SKOSMatcher tool developed by the European Union Joint Research Centre (JRC)[17] as part of the GENESIS project. This tool provides a user-friendly interface to allow users to manually define mapping properties between pairs of concepts from different ontologies (in this case, GEMET and the SBAs and other ontologies that may be added and aligned by the users as described in Section 4.4). From these definitions, the tool creates new SKOS alignment files. The present tool does not evaluate the consistency of the overall mapping. What is more, the tool does not do anything in case of cyclic path in alignments. This means that it is totally up to the users themselves to decide which concepts to align and which relations to use. However they are guided by the semantic definitions that each concept provides and by the already existing relations between concepts. The mapping will give the users an idea of the logics behind the overall alignment so as to be able to align other concepts that are coherent to this.

[16] SKOS W3C specifications mapping: www.w3.org/TR/SKOS-reference/#mapping

[17] EU-JRC: http://ec.europa.eu/dgs/jrc/index.cfm

4.4 Extending the Architecture with Additional Alignments

The alignment of the core ontologies aimed to provide a framework for users to annotate the resources to be used in the EuroGEOSS project, and also to undertake basic ontology-based query. However, it was recognized that there may be cases in a geospatial earth observation system in which users would need concepts that related to quite specialized scientific areas. For example, the EuroGEOSS project included the drought theme, and GEMET includes only two concepts relating to drought: drought and drought control, while the SBA categories and subcategories only include a subcategory: drought prediction. Drought scientists within the EuroGEOSS infrastructure required more detailed concepts to describe their resources. Thus a drought vocabulary is being developed and will be aligned to the core ontologies.

While the project resources were not sufficient to allow a full range of ontologies to be incorporated as core ontologies, the approach to ontology management was designed to be extendable (as in this drought example) to other ontologies that could be used for resource annotation and querying. To this end, users are provided with access to the SKOSMatcher tool, and also a set of detailed instructions to enable them to align their ontologies, including the following requirements.

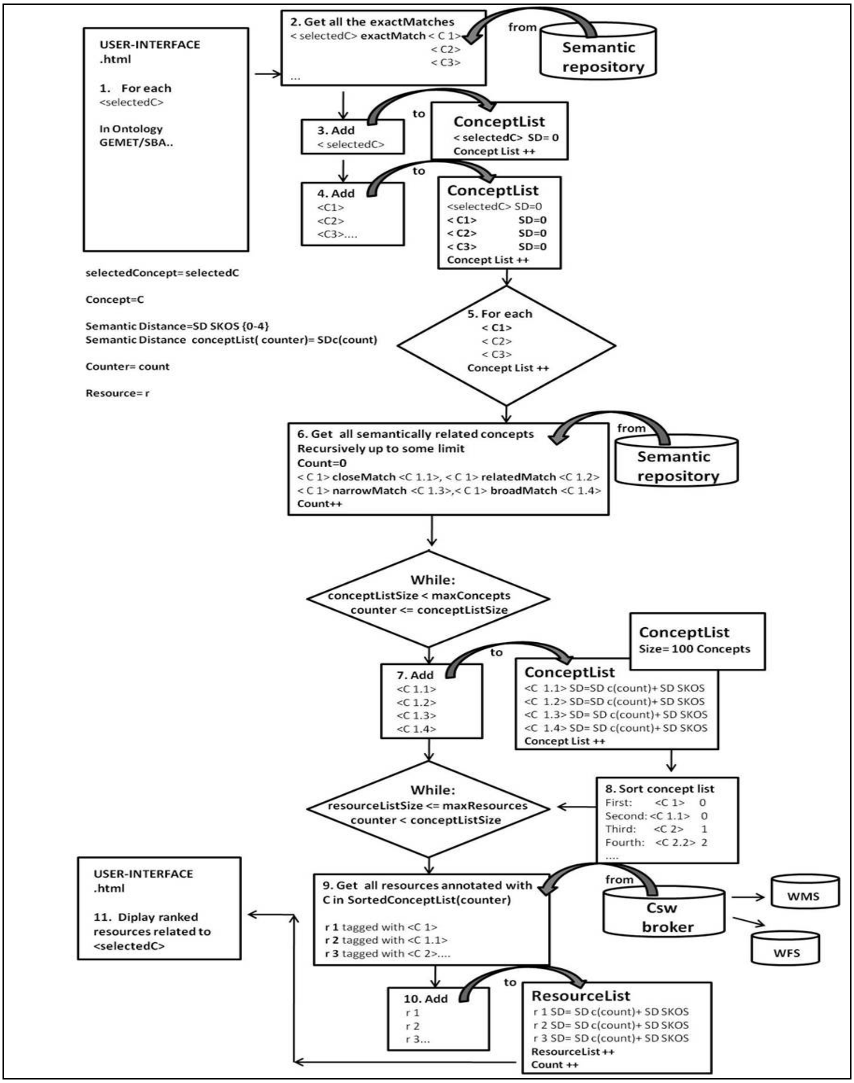
Firstly, if a user chose to align his/her own, additional ontology, s/he is required to align it with all existing ontologies. It is not sufficient to align with only one other ontology (for example, GEMET), but binary alignments with all ontologies already part of the framework are necessary. This is because, of the five mapping properties provided by SKOS, only the exactMatch property is transitive. Thus the only relations that could be assumed between any two existing aligned ontologies would be exactMatch relations, and these are likely to be in the minority because exact semantic equivalence between concepts in different ontologies is rare. For example, if a new ontology on Drought has a concept ‘drought severity’, which is a relatedMatch with a concept ‘drought’ in the existing aligned ontology GEMET, and there is also a concept ‘drought prediction’ in the vocabulary SBAs, which is aligned with GEMET ‘drought’ as a relatedMatch; this does not tell us anything about the relationship between concept ‘drought severity’ and ‘drought prediction’ except that they are both related to ‘drought’ in GEMET. This is not sufficient to support ontology based-resource discovery in an earth observation system.

Secondly, if a user chose to align an additional ontology, the user is required to align the entire ontology. It may be that a user is only interested in some portion of the ontology (a particular branch), to be aligned, and this would not necessarily be a problem from the point of querying (anything unaligned would simply not be returned with the query). However, it could cause problems for resource annotation and for subsequent mappings because after the alignment, resource providers may annotate their new resources with concepts that are from the unaligned portion of the ontology, and these will consequently not be returned by a query. It is also a potential problem for subsequent alignments because new ontologies that are aligned to the one partially aligned ontology may establish relations with concepts from the unaligned portion of the partially aligned ontology. This could lead to incomplete query results being returned. If a user wants to only align part of an ontology (which would only happen if there was a discrete branch for instance), the recommended course of action is for him or her to create a new ‘partial ontology’ that only included the branch of interest, and to publish it as a partial ontology to which other ontologies could align in full awareness of its partial status.

4.5 Using the Aligned Ontologies to Support Semantic Query

This Section illustrates the querying algorithm (Algorithm 1).

Algorithm 1. Semantic Query and Ranking Diagram


USER-INTERFACE
.html
1. For each
<selectedC>
In Ontology
GEMET/SBA..
2. Get all the exactMatches
< selectedC> exactMatch < C 1>
< C2>
< C3>
...
from
Semantic repository
3. Add
< selectedC>
to
ConceptList
< selectedC> SD= 0
Concept List ++
4. Add
<C1>
<C2>
<C3>....
to
ConceptList
<selectedC> SD=0
< C1> SD=0
< C2> SD=0
< C3> SD=0
Concept List ++
selectedConcept= selectedC
Concept=C
Semantic Distance=SD SKOS {0-4}
Semantic Distance conceptList(counter)= SDc(count)
Counter= count
Resource= r
5. For each
< C1>
< C2>
< C3>
Concept List ++
6. Get all semantically related concepts
Recursively up to some limit
Count=0
< C 1> closeMatch <C 1.1>, < C 1> relatedMatch <C 1.2>
< C 1> narrowMatch <C 1.3>,< C 1> broadMatch <C 1.4>
Count++
from
Semantic repository
While:
conceptListSize < maxConcepts
counter <= conceptListSize
ConceptList
Size= 100 Concepts
7. Add
<C 1.1>
<C 1.2>
<C 1.3>
<C 1.4>
to
ConceptList
<C 1.1> SD=SD c(count)+ SD SKOS
<C 1.2> SD=SD c(count)+ SD SKOS
<C 1.3> SD= SD c(count)+ SD SKOS
<C 1.4> SD= SD c(count)+ SD SKOS
Concept List ++
While:
resourceListSize <= maxResources
counter < conceptListSize
8. Sort concept list
First: <C 1> 0
Second: <C 1.1> 0
Third: <C 2> 1
Fourth: <C 2.2> 2
....
USER-INTERFACE
.html
11. Diplay ranked
resources related to
<selectedC>
9. Get all resources annotated with
C in SortedConceptList(counter)
r 1 tagged with <C 1>
r 2 tagged with <C 1.1>
r 3 tagged with <C 2>....
from
Csw broker
WMS
WFS
10. Add
r 1
r 2
r 3...
to
ResourceList
r 1 SD= SD c(count)+ SD SKOS
r 2 SD= SD c(count)+ SD SKOS
r 3 SD= SD c(count)+ SD SKOS
ResourceList ++
Count ++

Algorithm 1. Semantic Query and Ranking Pseudo-Code

```
‘get the initially selected concept and those that are exactly semantically equivalent
Add selectedConcept to conceptList with semantic distance 0 and increment conceptListSize
For each aligned ontology o
    For each concept c in o that has an ‘exactMatch’ relationship with selectedConcept
        Add c to conceptList with semantic distance 0
        Increment conceptListSize
‘go through each concept in conceptList and get all the semantically related concepts, recursively, up to some limit
counter = 0
‘while conceptList is not yet at its maximum size and there are still more concepts left in conceptList
While conceptListSize < maxConcepts and counter <= conceptListSize
    For each aligned ontology o
        For each concept c in o that has a defined SKOS relationship with conceptList(counter)
            Add c to conceptList with semantic distance = semantic distance of conceptList(counter) + semantic relation distance for the SKOS relationship between conceptList(counter) and c
            Increment conceptListSize
    ‘move to the next concept in conceptList and examine its relationships
    Increment counter
 ‘go through the concepts in conceptList and sort them in order of semantic distance using an existing sort algorithm
sortedConceptList = Sort(conceptList)
‘go through the sorted list and add related resources, until maxResources is reached
counter = 0
while counter < conceptListSize and resourceListSize <= maxResources
    Find each resource r that is annotated with the concept sortedConceptList(counter)
        If r is not already in resourceList
            Add r to resourceList
            Increment resourceListSize
    Increment counter
‘display them in ranked order
Display each r in resourceList in order
```

The algorithm was used to select and rank appropriate resources in response to a user query, making use of the SKOS relations both within and across aligned ontologies. The algorithm is not intended to be optimized for performance (this is being considered as part of the physical implementation), but instead to illustrate the approach to users as clearly as possible.

As a precursor to querying, it is assumed that resource providers (for example, agencies who provide a web service with a particular data set) have already annotated their resources with at least one ontology concept from any of the core or aligned ontologies. The algorithm uses these annotations to identify initial candidates for selection, and then uses the semantic relations and mapping properties to identify recursively other semantically related concepts and then the resources that are tagged with those related concepts. The results are ranked using a rating of semantic similarity. Thus it is not important which ontology or language to annotate the resource is chosen, because resources annotated with related concepts should also be identified. In the EuroGEOSS project, the CatMDEdit metadata editing tool [24] was used for semantic annotation.

The user query begins with the selection of a concept, again from any aligned ontology, after which Algorithm 1 is applied. The algorithm uses a very simple method for calculating semantic similarity that is based on semantic distance determined by the types of relationship that connect the concepts concerned.

A large number of alternative semantic matching methods are available and could also be employed here (with some adaptations) if required (for example, [25][26] [9] [28]). However, the approach described here concerns the use of the algorithm together with SKOS mapping relations among terms across ontologies to define numerically the extent of similarity they bear with each other.

Table 1 shows the semantic relation distances used. These were applied in an incremental fashion. For example, if a selected concept A is connected to concept B with a closeMatch relation, B has a semantic distance of 1 from A. If concept C is then connected to concept B with a broadMatch relation, concept C has a semantic distance of 4 from B and 5 from A. More complex semantic similarity algorithms could be used, and the semantic relation distances could be varied.

Table 1. Semantic Relation Distances

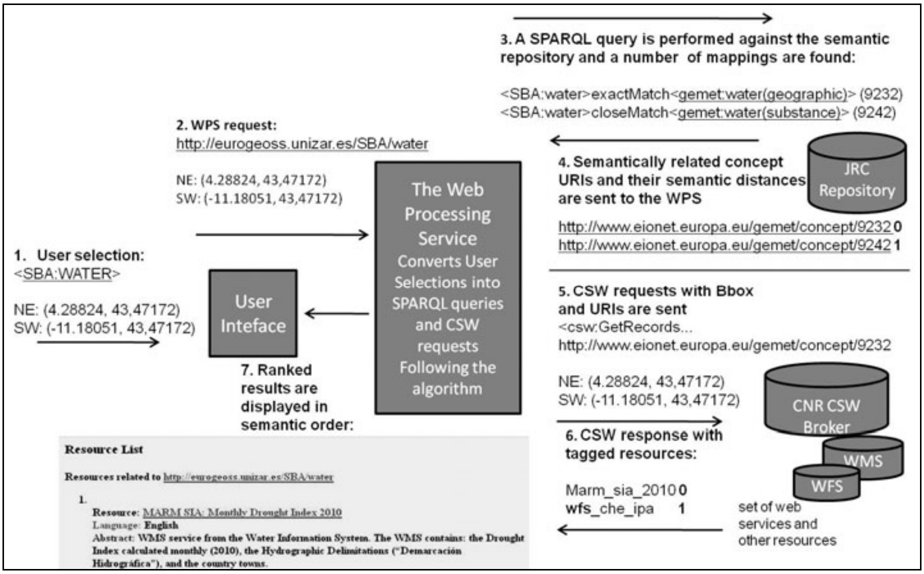
SKOS Mapping Property (between ontologies)	**SKOS semantic relation (within ontologies)**	**Semantic relation distance**
broadMatch	broader, broaderTransitive	4
narrowMatch	narrower, narrowerTransitive	3
relatedMatch	related	2
closeMatch		1
exactMatch		0

The role of the semantic relation distances, shown in Table 1, is crucial for the overall approach to assist the users in quickly finding the most suitable resources ranked so as they accurately match a query in order of semantic distance. The ranking is based on an assessment of how useful resources with the particular relations would be likely to be. For example, resources connected with an exactMatch relation to the selected concept are just as likely to be useful as if they were connected to the selected relation itself. Resources connected by a narrowMatch are thought to be more likely to be useful than those connected by a broadMatch relation, simply because they represent a specialization and thus a semantic refinement, so are likely to meet the requirements of the selected concept, while those connected by a broadMatch could potentially be so general as to not be at all useful. To give a real example, a concept ‘Forestry’ may have a narrower relation with ‘Deciduous Forests’ and a broader relation with ‘Agriculture, Fisheries and Forestry’. It could be claimed that resources tagged with ‘Deciduous Forests’ are more likely to be useful than those tagged with ‘Agriculture, Fisheries and Forestry’. However, there is certainly room for argument about these semantic relation distances, and they could be adjusted to suit the application concerned.

The size limits referred to in Algorithm 1 are mainly for performance reasons. The interoperable architecture discussed in this paper marshals multiple heterogeneous resources from around Europe using OGC web services. Therefore it is impractical to return a large number of resources, and a limit of perhaps 100 is suggested. If low numbers of resources are to be returned, it is likely that fewer concepts would need to be examined to yield the required number of resources, again improving performance. On the other hand, the greater the number of aligned ontologies included in the system, the more semantically similar concepts there are likely to be. Thus it may be appropriate to create a maximum number of concepts that is a function of the number of aligned ontologies, to yield the most semantically similar resources.

4.6 The Interoperable Architecture

The proposed method fits within an interoperable architecture. This enables multiple resource providers to provide their data and scientific models in the form of standards-based web services, and allows portals to be developed that can serve these heterogeneous resources through a single user interface.



1. User selection:
<SBA:WATER>

NE: (4.28824, 43,47172)
SW: (-11.18051, 43,47172)

2. WPS request:
http://eurogeoss.unizar.es/SBA/water

NE: (4.28824, 43,47172)
SW: (-11.18051, 43,47172)

3. A SPARQL query is performed against the semantic repository and a number of mappings are found:

<SBA:water>exactMatch<gemet:water(geographic)> (9232)
<SBA:water>closeMatch<gemet:water(substance)> (9242)

4. Semantically related concept URIs and their semantic distances are sent to the WPS

http://www.eionet.europa.eu/gemet/concept/9232 0
http://www.eionet.europa.eu/gemet/concept/9242 1

5. CSW requests with Bbox and URIs are sent
<csw:GetRecords...
http://www.eionet.europa.eu/gemet/concept/9232

NE: (4.28824, 43,47172)
SW: (-11.18051, 43,47172)

6. CSW response with tagged resources:

Marm_sia_2010 0
wfs_che_ipa 1

7. Ranked results are displayed in semantic order:

Resource List

Resources related to http://eurogeoss.unizar.es/SBA/water

1.
Resource: MARM SIA: Monthly Drought Index 2010
Language: **English**
Abstract: WMS service from the Water Information System. The WMS contains: the Drought Index calculated monthly (2010), the Hydrographic Delimitations ("Demarcación Hidrográfica"), and the country towns.

Fig. 1. Illustrative High Level Architecture of the WPS

Figure 1 illustrates the high level architecture. The architecture accesses two sources of information using open standards. These two sources of information are: a) The ontology repository (administered by the JRC) contains the ontologies and the alignments within and between the ontologies, in SKOS format. This is used to search for semantically similar concepts in any language, so that all relevant resources may be retrieved in response to a query, even if they are annotated in another language, using a different ontology term; b) The CSW broker[18], administered by the Italian National Research Council (CNR),[19] mainly provides access to metadata for other resources through distributed thematic registries.

Examples of retrievable services (resources) include geographic information datasets and map services providing geographic visualizations on the three thematic areas of biodiversity, forestry and drought, gazetteers, feature services etc.

The architecture makes use of web service standard specifications from the Open Geospatial Consortium (OGC) widely used throughout the international geospatial community. Specifically, the specifications employed are: a) The Web Processing Service (WPS) Specification [29] is used as a wrapper around the process that takes a simple user request, executes Algorithm 1. It creates SPARQL queries to retrieve ontology terms and their SKOS mappings, and CSW (see below) requests to retrieve resource metadata to present to the user. WPS defines a request and response interface to ensure interoperability; b) The ISO 19115/119 Application Profile for CSW [30] (which forms part of the Catalogue Services for the Web specification) is used as an interface to the CSW broker [31]; c) The Web Map Service [32], Web Feature Service [33], and Web Coverage Service [34].

The aim of this architecture is to support the use of multiple multilingual ontologies by implementing Algorithm 1, but also to provide an interoperable, distributed solution that adopts international standards.

5 Discussion

The approach described in Section 4 was implemented as part of the EuroGEOSS project. Thematic international experts were instructed on how to align their own ontologies. The semantic querying approach was implemented using the architecture described. This Section describes issues that arose in this process and suggests possible solutions.

5.1 Transitivity of SKOS Relations

Possible ambiguities still need to be solved on the transitivity of the SKOS mapping relations, especially with multiple ontologies. Specifically, the broadMatch and narrowMatch relations are defined as being non-transitive, since only the SKOS exactMatch mapping property (for use across, rather than within ontologies) is transitive. This means that if the term ‘impacts of humans on water cycle’ in GEMET is broader than the term ‘water pollution’ in the SBAs that is broader than the term ‘sea pollution’ in a new ontology, this does not entail that the term ‘impacts of humans on water cycle’ in GEMET is broader than ‘sea pollution’ in a new ontology (this would need to explicitly aligned using the broadMatch relation). This state of affairs is one of the reasons for the requirement for a manual alignment with multiple ontologies in this architecture. Broad transitive and narrow transitive relations are available in SKOS for use within ontologies, but not as mapping properties between

[18] CSW broker with a list of federated resources: `http://217.108.210.73/broker/`

[19] CNR: `http://www.cnr.it/sitocnr/Englishversion/Englishversion.html`

ontologies. SKOS (particularly between ontologies) is generally a simple representation system for non experts to use, and a more complex set of relations would perhaps dissuade users from contributing.

5.2 Alignment of GEMET and the GEOSS SBAs

One of the biggest challenges while aligning the core ontologies derived from the very structure of the ontologies and from some limitations of the SKOS system. For example in the poly-hierarchical structure of the GEMET Thesaurus, Themes and Groups have been defined as collections (grouping of concepts) and their concepts.

Unfortunately though, the SKOS system does not provide an approach to the modelling of relations among different entities (for example, concepts with collections) across independent ontologies as underlined also in [35]. To put it another way, SKOS does not provide a method for relating a concept in one ontology with a group or a theme term in another (Fig. 2 below). Possible solutions include the development of a different classification reorganization within the SKOS model, to allow mapping properties to be defined among collections as well as concepts.

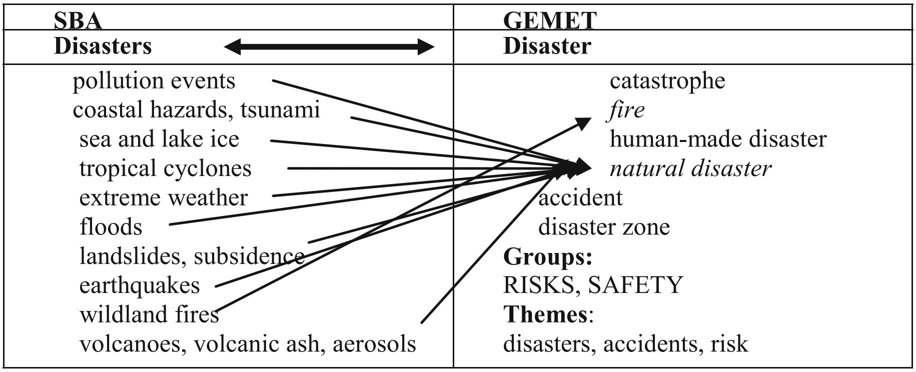


Fig. 2. SBA-GEMET branch-mapping. The *thinner arrows* pointing from left to right represent *SKOS broadMatch*; the *thicker double arrow* pointing at both sides represents SKOS *exactMatch*. The graph follows the SKOS notation, so if in <floods> BroadMatch <natural disaster> the latter is broader, the arrow points towards the broader term.

5.3 Ontology Versioning

During the course of the project, the multi-lingual GEOSS SBA categories and subcategories were being updated due to recent amendments in translation. This made the process of alignment difficult since the versions being used were unstable, but did not affect the technical implementation, because concepts were retrieved through URIs, which are linguistically independent for the machine recognition. This means that the implementation of new languages could still be a work-in-progress without invalidating severely the implementation of the approach, as long as the semantic alignment work was completed.

5.4 Multilingual Mapping Limitations

The architecture described in this paper assumes that single, multilingual ontologies are used, in which a single ontology concept has multiple translations, expressed in SKOS as alternative labels referred by a single URI. However, in reality it is often the case that a concept in one language does not have an exact semantic match with another language. Therefore, a more complete multilingual solution would create separate ontologies in each language, and define the relationships between the concepts from different languages using SKOS mapping properties. Algorithm 1 could easily be adapted to this ontology architecture, but additional effort in ontology alignment would be required. Exploration of this more advanced multilingual architecture is anticipated in the future.

5.5 User Evaluation

The approach taken so far has been implemented as described and initial evaluation from testing has been found to be positive. In the future we intend to evaluate the approach further once an additional drought ontology has been added.

What is more, shortly we plan to evaluate the interface with users to determine their appreciation of the multilingual, semantic query support, using the Microsoft Desirability Toolkit[20] and unstructured interview questions.

6 Conclusions

This paper has presented an approach to the management of the semantic aspects of information in a large, heterogeneous geospatial system in a multi-theme, multilingual context. In particular, it has described work carried out involving the implementation of an integrated semantic engine to support such a system based on mappings between core ontologies and more specialized ontologies.

The approach is extensible in including ontologies that cover particular thematic areas of interest, if they are not adequately covered by the core ontologies. However, performance impacts would be expected as additional ontologies are included.

The approach is also multi-lingual, as the SKOS defines multilingual versions of a particular concept, and the very algorithm queries terms in a range of different languages. The user may thus be presented with services in different languages in response to his or her query.

Future work by the authors is extending the existing approach by adding simple natural language querying tools, particularly addressing spatial relations between the concepts that are selected from the ontologies. This natural language work is also particularly focused on a multi-lingual environment in which the different ways of expressing spatial relations in different languages (including non Indo-European languages) are accommodated.

Acknowledgements. The project described in this paper was funded by the European Union, and their support is gratefully acknowledged as are the helpful comments of the reviewers.

[20] www.microsoft.com/usability/UEPostings/DesirabilityToolkit.doc

References

1. Lassila, O., McGuinness, D.: The Role of Frame-Based Representation on the Semantic Web. KSL Tech. Report Number KSL-01-02 (2001)
2. Hyvönen, E.: Preventing ontology interoperability problems instead of solving them. Semantic Web 1(1-2), 33–37 (2010), doi:10.3233/SW-2010-0014
3. Noy, N.F., Musen, M.: Anchor-PROMPT: Using Non-Local Context for Semantic Matching. In: Workshop on Ontologies and Information Sharing at IJCAI, Stanford (2001)
4. Noy, N.F., Musen, M.: SMART: Automated Support for Ontology Merging and Alignment (1999)
5. Noy, N.F., Klein, M.: Ontology Evolution: Not the Same as Schema Evolution. In: Knowledge and Information Systems, vol. 6, pp. 428–440. Springer-Verlag London Ltd., Heidelberg (2004)
6. Guarino, N.: Formal Ontology in Information Systems. In: Proceedings of FOIS 1998, Trento, Italy, pp. 3–15. IOS Press, Amsterdam (1998)
7. Cruz, I.F., Sunna, W., Makar, N., Bathala, S.: A visual tool for ontology alignment to enable geospatial interoperability. In: Web Semantics: Science, Services and Agents on the World Wide Web, vol. 5, pp. 39–49 (2007)
8. Noy, N.F., Musen, M.: Anchor-PROMPT: Using Non-Local Context for Semantic Matching. In: Workshop on Ontologies and Information Sharing at IJCAI (2001)
9. Rodríguez, M.A., Egenhofer, M.J.: Determining Semantic Similarity among Entity Classes from Different Ontologies. IEEE Transactions on Knowledge and Data Engineering 15(2), 442–456 (2003)
10. Lan, G., Huang, Q.: Ontology-based Method for Geospatial Web Services Discovery. In: Advances in Intelligent Systems Research, International Conference on Intelligent Systems and Knowledge Engineering, ISKE 2007 (2007)
11. Gangemi, A., Pisanelli, D.M., Steve, G.: An Overview of the ONIONS Project: Applying Ontologies to the Integration of Medical Terminologies. Data Knowledge Engineering 31(2), 183–220 (1999)
12. Stumme, G., Madche, A.: FCA-Merge: Bottom-up merging of ontologies. In: 7th Intl. Conf. on Artificial Intelligence (IJCAI 2001), pp. 225–230 (2001)
13. Mitra, P., Wiederhold, G., Kersten, M.L.: A graph-oriented model for articulation of ontology interdependencies. In: Zaniolo, C., Grust, T., Scholl, M.H., Lockemann, P.C. (eds.) EDBT 2000. LNCS, vol. 1777, pp. 86–100. Springer, Heidelberg (2000)
14. Sunna, W., Cruz, I.F.: Structural Alignment Methods with Applications to Geospatial Ontologies. In: Janowicz, K., Raubal, M., Schwering, A., Kuhn, W. (eds.) Transactions in GIS, Special Issue on Semantic Similarity Measurement and Geospatial Applications, vol. 12(6), pp. 683–711 (2008)
15. Sunna, W., Cruz, I.F.: Structure-Based Methods to Enhance Geospatial Ontology Alignment. In: Fonseca, F., Rodríguez, M.A., Levashkin, S. (eds.) GeoS 2007. LNCS, vol. 4853, pp. 82–97. Springer, Heidelberg (2007)
16. Grütter, R., Bauer-Messmer, B., Frehner, M.: First Experiences with an Ontology-Based Search for Environmental Data. In: 11th AGILE International Conference on Geographic Information Science 2008, pp. 1–9. University of Girona, Spain (2008)
17. Baglioni, M., Giovannetti, E., Masserotti, M.V., Renso, C., Spinsanti, L.: Ontology-supported Querying of Geographical Databases. Transactions in GIS 12(s1), 34–44 (2008)
18. Klien, E., Lutz, M., Kuhn, W.: Ontology-Based Discovery of Geographic Information Services: An Application in Disaster Management. Computers, Environment and Urban Systems 30, 102–123 (2006)

19. Lacasta, J., Nogueras-Iso, J., Béjar, R., Muro-Medrano, P.R., Zarazaga-Soria, F.J.: A Web Ontology Service to Facilitate Interoperability within a Spatial Data Infrastructure: Applicability to Discovery. Data & Knowledge Engineering 63, 947–971 (2007)
20. Vidal, V.M.P., Sacramento, E.R., de Macêdo, J.A.F., Casanova, M.A.: An Ontology-Based Framework for Geographic Data Integration. In: Heuser, C.A., Pernul, G. (eds.) ER 2009. LNCS, vol. 5833, pp. 337–346. Springer, Heidelberg (2009)
21. Zhan, Q., Zhang, X., Lic, D.: Ontology-Based Semantic Description Model for Discovery and Retrieval of Geospatial Information. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 32(Part B4) (2008)
22. Nida, E.A.: Componential Analysis of Meaning: an Introduction to Semantic Structures. Mouton, The Hague (1975)
23. Stock, K.: The Representation of Geographic Object Semantics Using Inclusion Rules, PhD thesis, GIS/LIS 1998 Annual Conference and Exposition held at Fort Worth, Texas, November 8-12 (1998)
24. Nogueras-Iso, J., Barrera, J., Gracia-Crespo, F., Laiglesia, S., Muro-Medrano, P.R.: Integrating catalog and GIS tools: access to resources from CatMDEDit thanks to gvSIG. In: 4as Jornadas Internacionales gvSIG, Valencia, Spain, December 3-5, pp. 1–10 (2008), Online resource at http://catmdedit.sourceforge.net/
25. Ge, J., Qiu, Y.: Concept Similarity Matching Based on Semantic Distance. In: Fourth International Conference on Semantics, Knowledge and Grid (2008)
26. Hau, J., Lee, W., Darlington, J.: A Semantic Similarity Measure for Semantic Web Services, Chiba, Japan, May 10-14 (2005)
27. Whiteside, A., Evans, J.D.: OGC® Web Coverage Service (WCS) Implementation Standard, OGC 07-067r5 Version: 1.1.2 (2008)
28. Schwering, A., Raubal, M.: Spatial Relations for Semantic Similarity Measurement. In: Akoka, J., et al. (eds.) ER Workshops 2005. LNCS, vol. 3770, pp. 259–269. Springer, Heidelberg (2005)
29. Schut, P.: OpenGIS® Web Processing Service, OGC 05-007r7. OpenGIS® Standard, version 1.0.0 (2007)
30. Voges, U., Senkler, K.: OpenGIS® Catalogue Services Specification 2.0- ISO 19115/119 Application Profile for CSW, OGC 07-006r1, Version 2.0.2 (2005); Panzer, M., Lei Zeng, M.: Modeling Classification Systems in SKOS: Some Challenges and Best-Practice Recommendations. In: Proceedings of the International Conference on Dublin Core and Metadata Applications, DC-2009–Seoul (2009)
31. Nebert, D., Whiteside, A., Vretanos, P.: OpenGIS® Catalogue Services Specification, OGC 07-006r1, Version 2.0.2 (2007)
32. De la Beaujardiere, J.: OGC Web Map Service Interface, OGC 03-109r1, Version: 1.3.0 (2004)
33. Vretanos, P.: OpenGIS® Web Feature Service Implementation Specification, OGC 04-094, Version: 1.1.0 (2005)
34. Whiteside, A., Evans, J.D.: OGC® Implementation Standard, OGC 07-067r5, Version 1.1.2, 2nd release (2008)
35. Panzer, M., Lei Zeng, M.: Modeling Classification Systems in SKOS: Some Challenges and Best-Practice Recommendations. In: Proceedings of the International Conference on Dublin Core and Metadata Applications, DC-2009–Seoul (2009)

Identifying Geographical Processes from Time-Stamped Data

Claudio E.C. Campelo, Brandon Bennett, and Vania Dimitrova

School of Computing,
University of Leeds,
Leeds,
LS2 9JT, UK
{sccec,b.bennett,v.g.dimitrova}@leeds.ac.uk

Abstract. Humans tend to interpret a temporal series of geographical spatial data in terms of *geographical processes*. They also often ascribe certain properties to processes (e.g. a process may be said to *accelerate*). Given a spatial region of observation, distinct properties may be observed in different subregions and at different times, which causes difficulties for humans to identify them. The conceptualisation of geographical features and their correlation with geographical phenomena may provide a human like approach to analyse large spatio-temporal datasets. This paper presents a representational model and a reasoning mechanism to analyse evolving geographical features and their relationship to geographical processes. The proposed approach comprises methods of relating occurrences of geographical events to geographical processes which is said to proceed over time. We introduce an initial set of properties which can be associated with several geographical processes. We consider this as a first step towards a more general model for representing and reasoning about geographical processes.

Keywords: Spatial Reasoning, Temporal Reasoning, Ontology, Geographical Processes.

1 Introduction

Geographical features may change over time. These changes sometimes occur continuously, such as clouds moving in the atmosphere, and sometimes happen in cycles, as for example the seasonal variation in vegetation. Some authors (e.g. [12,14,15]) have classified this type of change as *spatio-temporal processes*.

The features involved in such processes may undergo a variety of spatial transformations. For example, they can grow, shrink or move. When a spatio-temporal process comprises only changes in geographical features, we conceive it as a *geographical process*, and the geographical features involved are taken as being the *participants* in the process. Examples of geographical processes are *deforestation*, *urbanisation* and *desertification*. The former, for instance, may be defined in terms of changes in a feature of type *forest*. In addition, these processes are

C. Claramunt, S. Levashkin, and M. Bertolotto (Eds.): GeoS 2011, LNCS 6631, pp. 70–87, 2011.

generally associated with a set of *properties* which may be ascribed to them (for example, a process may be described as being *constant*, or *intermittent*, or *slowing down*, or *accelerating*). We introduce a set of properties which can be applied to a variety of geographical processes: *initiation* and *cessation*, *acceleration*, *deceleration* and *constant proceeding.* We also present an approach to identify whether a process is proceeding on the basis of given geographic data.

In this paper, we describe a representational model for those topographical and mereological changes in the geographical space, which we characterise as a geographical process. We also describe a reasoning mechanism to identify features whose changes are associated with properties of a geographical process. In order to evaluate our conceptual model, we have developed a system prototype which links an ontology to a spatio-temporal dataset. The prototype takes a temporal series of spatial data as an input and returns a set of features which matches a user query, which consists of the process' properties to be identified and spatial and temporal constraints.

Defining an appropriate representation for geographical processes requires dealing with issues regarding their relationship to events and objects. While the former may be conceived as constituents of processes [12], the latter may be regarded as *participants* in processes. In addition, objects are associated with a spatial extent at any one time and persist through changes in their attributes [11]. Other important issues are how to define the relation between process types and particular process instances, and how to associate specific spatial and temporal boundaries with process instances. Accordingly, our semantic model comprises methods of defining types and instances and provides an approach to define such spatial and temporal boundaries.

The structure of the paper is as follows. The next section discusses the related work. Section 3 presents the motivation and research challenges. This is followed by the discussion about the approach to representing timestamped data in Section 4. Then Section 5 describes a logical framework comprising formal descriptions of space, time, events, processes, geographical objects and their related aspects. In Section 6 we introduce the semantic definition of a geographical process. Then Section 7 presents formal definitions for the set of process properties which we have proposed. Finally, Section 8 draws some conclusions and discusses directions for further work.

2 Related Work

Processes have been investigated in different areas, such as Philosophy, Linguistics, Business and in several sub-areas of Computer Science. In Geographical Information Systems (GIS), different approaches have been developed to deal with processes, and an assorted terminology has been applied (e.g. *geo-processes*, *geo-phenomena*, *dynamic GIS*, *spatio-temporal GIS*). In the field of Knowledge Representation, *spatio-temporal reasoning* [5,10,24] and *reasoning about spatio-temporal changes* [16] have been investigated. Theories involving *objects*, *events*, *states* and *process* have also become of interest [12,15].

Some research in GIS has been presented as an approach to handling real-world phenomena, which appears in the literature under a variety of different names, such as *process models*, *reality-representation systems* and *modelling-systems*. Nonetheless, the development of such systems has been limited to a particular area of application (e.g. meteorology, traffic studies, population studies) and usually with the purpose of simulation and prediction. Examples can be found in [18,23].

Modelling approaches for dynamic geospatial domains based on the concepts of objects and events have been studied in GIS. Worboys and Hornsby [25] discuss how such models extend traditional object-based geospatial models. Approaches to modelling spatio-temporal process based on object-oriented data models have also been proposed [3]. Claramunt and Thériault present a taxonomy of processes and semantics for modelling spatio-temporal evolution within GIS, based on a event-oriented representation of spatial dynamics [4].

Galton [11] argues that a spatio-temporal geo-ontology must comprise appropriate forms of representation to do justice to both the *field-based* and *object-based* views of the world [13], extended appropriately to consider the temporal dimension. Additionally, it should provide different views of spatio-temporal extents, especially with reference to phenomena such as storms, floods and wildfires. Field-based approaches to simulating geographical processes have been proposed by using *cellular automata*. Examples can be found for simulation of wildfires [2,17] and urban spreading [1,23]. On the other hand, *agent-based models* have been suggested as an object-based approach to handling geographical processes. These models have been proposed, for example, to simulating landscape evolution [21], urban development [7] and traffic systems [9].

Some authors have also presented ontological approaches to representing geographical processes. For instance, Devaraju and Kuhn [8] present an ontology to represent relations between geographical processes and observed properties originated from Geo-Sensor Networks, allowing the representation of a process and its participants (i.e. physical objects and substances). However, this is distinct from the approach presented here, since it is mainly concerned with modelling geographical processes in terms of the interaction between their participants and the physical and chemical transformations involved in their execution, whereas this work draws more attention to the identification of geographical processes from spatial changes observed in geographical features over time.

3 Motivation and Research Challenges

A temporal series of spatial data representing evolving geographical features is often interpreted by humans in terms of geographical processes and their associated properties. Given a spatial region and a time interval of observation, distinct manifestations of a geographical process may be identified simultaneously in different sub-regions, or at different sub-intervals in the same sub-region. For example, a desertification process may accelerate in a subregion during a certain period of time, whilst another process of the same type proceeds steadily in another subregion during the same period. When dealing with reasonably large datasets, this dynamicity causes difficulties for humans to identify some manifestations of geographical processes. Therefore, the conceptualisation of dynamic geographical features and their correlation with geographical processes may provide an approach that can augment human's ability to analyse large spatio-temporal datasets.

Remote Sensing is an evolving research area which collects useful data about the physical world. A considerable amount of data produced is the result of digital processing of satellite imagery and aerial photography. These areas of research are mostly concerned with the identification and classification of different features that compose the geographical space, and frequently generate vector spatial data as an output, such as described in [20] and [22]. Moreover, temporal series of spatial data have been generated by producing images of a given region at different times. However, since this field of study is reasonably recent, many spatio-temporal datasets are still being produced and are not fully exploited in GIS. Therefore, this increasing availability of data reinforce the demand for efficient approaches to link such spatio-temporal data to a reasoning mechanism in order to investigate geographical processes.

The capability of defining precise instantiations of a process (both in space and in time) is a fundamental issue to provide a suitable approach to reasoning about its properties. A representational model which takes into account those questions needs to be flexible in consideration of different interpretations which may arise for some dynamic properties investigated. For example, to maintain that a deforestation process is accelerating in a given spatial region R and time interval I requires the definition of a 'deforestation expansion rate'. However, there may be distinct interpretations for this rate. For instance, it could relate the area of the portion of R deforested during I either to the previously existing deforested portion of R or to the total area of R. Also the rate could be measured in terms of absolute area or as a percentage.

Several approaches to reasoning about space and time and different methods of modelling geographical processes have been proposed, however, an approach to developing an ontology grounded upon a spatio-temporal data to reasoning about geographical processes and their spatio-temporal changing properties is still missing. Grounding an ontology upon the data requires work at multiple levels, both to select the appropriate set of predicates to be grounded and formulating a suitable representation for the data. In addition, a considerable amount of work on geometrical computation should be done to enable spatial predicates defined in the conceptual level to be interpreted at the data level.

4 Time-Stamped Data

As stated before, we propose a logical framework which can be linked explicitly to a spatial-temporal dataset in order to develop practical applications. Spatio-temporal datasets are distributed in a variety of formats, which include aspects of geometrical representation of space, representation of temporal elements, scale, granularity and other important aspects. This Section presents our approach to the representation and storage of data in such a way that reasoning about dynamic geographical features is possible.

We present some relations which hold between different forms of data representation, in order to provide a way to derive implicit data in reduced datasets. They are described in terms of definitions and axioms in first order logic, indexed by D and A, respectively. In axioms, free variables are implicitly universally quantified with maximal scope. We employ the Region Connection Calculus (RCC) [19] as the theory of *space*, using the following spatial relations between spatial regions: *overlaps* $\mathsf{O}(x, y)$, *externally connected* $\mathsf{EC}(x, y)$ and *part of* $\mathsf{P}(x, y)$.

We assume a total linear reflexive ordering on *time*, and use explicit time variables t_i. These variables can be compared by equality ($t_1 = t_2$) and ordering ($t_1 < t_2$) relations and can be quantified over in the usual way ($\forall t[\phi(t)]$). In addition, the predicates $\mathsf{Instant}(s)$ and $\mathsf{Interval}(s)$ are employed to distinguish instants and intervals. We also use the functions $\mathsf{b}(i)$ and $\mathsf{e}(i)$, which return an instant corresponding to the beginning and the end of an interval i, respectively.

The data is stored as factual elements asserted in a knowledge base. Storing the data in a logical fashion allows us to derive implicit data and provides natural way to link the logical framework to the knowledge base. The spatio-temporal data consist of attributed polygons which are associated with timestamps. *Attributes* describe either types of region coverage (e.g. *forested, arid*) or types of geographical features (e.g. *forest, desert*). *Polygons* represent spatial regions or geographical features. We are particularly interested in geographical features which can be modelled as the maximal well-connected regions[1] of some particular coverage, as shall be discussed later in this Section. *Timestamps* are used to represent both time instants and time intervals. This data is structured as follows:

$\mathcal{D} \subseteq A \times P \times S$, where:

A is the set of attributes;

P is the set of two-dimensional simple[2] polygons;

S is the set of time instants and intervals, defined as:

$S = \{ \langle t_1, t_2 \rangle \mid t_1, t_2 \in T \wedge t_1 \leq t_2 \}$, where T is the set of timestamps;

According to this definition, an instant is represented by a zero-length interval (i.e. when $t_1 = t_2$).

[1] The term ‘well-connected region’ is used here in agreement with the discussion and definitions given in [6].

[2] A *simple polygon* is one whose boundary does not cross itself. However, in order to represent spatial regions which can have holes, the set S may contain *weakly simple polygons*, in which some sides can ‘touch’ but cannot ‘cross over’. See `http://en.wikipedia.org/wiki/Simple_polygon` for further explanation and examples.

A datum is a tuple assuming the form $\langle a, p, s\rangle$, where p is a polygon, a is an attribute and s is an instant or interval. These data elements are stored as asserted facts using the predicate *Spatio-temporal Attributed Region* $\mathsf{Star}(a, p, s)$. For readability, we use in this paper the predicate $\mathsf{A\text{-}Star}(a, p, s)$ to indicate that the fact is *explicitly asserted* in the knowledge base, whereas the truth of $\mathsf{Star}(a, p, s)$ is determined by the semantics of attribute a and the geographic characteristics of the geo-referenced polygon p, whether or not it is actually asserted in the knowledge base. Consequently, we specify the axiom A1 to assure that $\mathsf{Star}(a, p, s)$ is true if the corresponding fact (explicitly asserted) is true.

A 1 $\mathsf{A\text{-}Star}(a, p, s) \rightarrow \mathsf{Star}(a, p, s)$

There are many different ways in which an attribute can be used to describe a spatial region with respect to a time point or interval. Since we treat an attribute a as a special kind of entity, we can use predicates to classify attributes and first-order formulae to axiomatise semantic characteristics and inter-dependencies of attributes.

Our data model currently supports a geographic knowledge base in which the following kinds of attribute are recorded:

- $\mathsf{CAtt\text{-}Hom}(x)$ — *homogeneous coverage attributes* are applied to denote spatial regions which are regarded as covered by a single type of coverage.
- $\mathsf{CAtt\text{-}Het}(x)$ — *heterogeneous coverage attributes* are employed to denote spatial regions which may contain multiple types of coverage.
- $\mathsf{FAtt\text{-}Sim}(x)$ — *simple feature attributes* are applied to denote geographical features which cannot be composed by other geographical features and that every region which is part of it must have the same coverage.
- $\mathsf{FAtt\text{-}Com}(x)$ — *compound feature attributes* are applied to denote geographical features which may be composed by other geographical features or that may contain regions with different coverages.

However, the semantic model (which shall be discussed later in this paper) currently focus on geographical processes which affect simple features. For this reason, this paper draws particular attention to the attributes $\mathsf{CAtt\text{-}Hom}(x)$ and $\mathsf{FAtt\text{-}Sim}(x)$. More general predicates for attribute types are defined in D1, D2 and D3: *coverage attribute* $\mathsf{CAtt}(x)$, *feature attribute* $\mathsf{FAtt}(x)$ and *attribute* $\mathsf{Att}(x)$, respectively.

D 1 $\mathsf{CAtt}(x) \equiv_{def} \mathsf{CAtt\text{-}Hom}(x) \vee \mathsf{CAtt\text{-}Het}(x)$

D 2 $\mathsf{FAtt}(x) \equiv_{def} \mathsf{FAtt\text{-}Sim}(x) \vee \mathsf{FAtt\text{-}Com}(x)$

D 3 $\mathsf{Att}(x) \equiv_{def} \mathsf{CAtt}(x) \vee \mathsf{FAtt}(x)$

We can now specify the axiom A2 which relates the predicate $\mathsf{Star}(a, p, s)$ to the elements to which it applies. In this axiom, we use the predicate $\mathsf{Polygon}(p)$ to assert that p is a two-dimensional simple polygon.

A 2 $\mathsf{Star}(a, p, s) \rightarrow (\mathsf{Polygon}(p) \wedge \mathsf{Att}(a) \wedge (\mathsf{Instant}(s) \vee \mathsf{Interval}(s)))$

A variety of geospatial information may be represented using *STAR* relations. However, we define axioms which restrict the data model to deal with a set of useful kinds of representations which are currently supported by our semantic model. When a *STAR* element is associated with a time instant, it represents a snapshot of a spatial region or geographical feature at this time point. On the other hand, when this element is associated with a time interval, it denotes either a spatial region whose coverage has been entirely changed during such interval or a geographical feature which has been created during the specified interval (and did not exist before the interval). These forms of representing data elements enable our model to use datasets distributed using different approaches to represent the data, i.e., containing different combinations of these types of data elements. Thus we shall discuss how these elements are related so that it is possible to deduce implicit data from limited datasets.

We employ the *non-reflexive* and *asymmetric* logical relation *Can be Part* $\mathsf{CP}(a_1, a_2)$ to specify the cases where part-hood relations can hold between STAR elements associated with different attributes. This relation is applied to associate: *homogeneous* and *heterogeneous coverage attributes*; or *homogeneous coverage attributes* and *simple feature attributes*; or *heterogeneous coverage attributes* and *compound feature attributes.* These relationships must be explicitly asserted as facts in the knowledge base. The first and second case ensures that a homogeneous region whose coverage type is denoted by a_1 can be either part of a heterogeneous region whose coverage type is denoted by a_2; or part of a simple feature whose type is denoted by the attribute a_2. The last case ensures that a heterogeneous region whose coverage is denoted by the attribute a_1 can be part of a compound feature whose type is denoted by the attribute a_2. The axioms A3 and A4 are specified to ensure the properties of this relation are preserved.

A 3 $$\begin{aligned}\mathsf{CP}(a_1, a_2) \rightarrow\ & (\mathsf{CAtt\text{-}Hom}(a_1) \wedge \mathsf{CAtt\text{-}Het}(a_2)) \vee \\ & (\mathsf{CAtt\text{-}Hom}(a_1) \wedge \mathsf{FAtt\text{-}Sim}(a_2)) \vee \\ & (\mathsf{CAtt\text{-}Het}(a_1) \wedge \mathsf{FAtt\text{-}Com}(a_2))\end{aligned}$$

A 4 $$\begin{aligned}&\mathsf{Star}(a_1, p_1, s) \wedge \mathsf{Star}(a_2, p_2, s) \wedge \\ &\quad (a_1 \neq a_2) \wedge \mathsf{P}(p_1, p_2) \rightarrow \mathsf{CP}(a_1, a_2)\end{aligned}$$

The axiom A5 restricts the cases where part-hood relations may hold involving STAR elements associated with the same attribute.

A 5 $\mathsf{Star}(a, p_1, s) \wedge \mathsf{Star}(a, p_2, s) \wedge \mathsf{PP}(p_1, p_2) \rightarrow \mathsf{CAtt}(a)$

The axioms A6, A7, and A8 are also specified to ensure that a *feature attribute* must be related to one (and only one) *coverage attribute*, and a *heterogeneous coverage attribute* must be related to at least one *homogeneous coverage attribute*.[3]

[3] Since *compound feature attributes* cannot be related to *homogeneous coverage attributes* directly using $\mathsf{CP}(a_1, a_2)$, such features are specified by relating a *compound feature attribute* to a *heterogeneous coverage attribute* which in turn is related to one or many *homogeneous coverage attributes.*

A 6 $\mathsf{FAtt}(a) \rightarrow \exists a'[\,\mathsf{CAtt}(a') \wedge \mathsf{CP}(a', a)\,]$

A 7 $\mathsf{CP}(a_1, a_2) \wedge \mathsf{FAtt}(a_2) \rightarrow \mathsf{CAtt}(a_1) \wedge \neg\exists a[\,(\mathsf{CP}(a, a_2) \wedge a \neq a_1)\,]$

A 8 $\mathsf{CAtt\text{-}Het}(a) \rightarrow \exists a'[\,\mathsf{CAtt\text{-}Hom}(a') \wedge \mathsf{CP}(a', a)\,]$

An additional axiom is specified for the case when the relation $\mathsf{CP}(a_1, a_2)$ is applied to associate *homogeneous coverage attributes* and *simple feature attributes*, in order to preserve the properties of homogeneity and simplicity assigned to these types of attributes, respectively.

A 9
$$(\mathsf{CP}(a_1, a_2) \wedge \mathsf{CAtt\text{-}Hom}(a_1) \wedge \mathsf{FAtt\text{-}Sim}(a_2)) \leftrightarrow \forall a[\,\exists p_1 p_2 s[\,\mathsf{Star}(a_1, p_1, s) \wedge \mathsf{Star}(a, p_2, s) \wedge \mathsf{P}(p_1, p_2) \wedge \mathsf{FAtt\text{-}Sim}(a)\,] \rightarrow a = a_2\,]$$

We have stated that many geographical features can be modelled as the maximal well-connected regions of some particular coverage. This is specified using the axiom A10.

A 10
$$\mathsf{A\text{-}Star}(a, p, s) \wedge \mathsf{FAtt}(a) \leftrightarrow \neg\exists p' a'[\,\mathsf{Star}(a', p', s) \wedge \mathsf{CAtt}(a') \wedge \mathsf{CP}(a', a) \wedge \mathsf{PP}(p, p')\,]$$

A geographical feature also denotes a spatial region with the same spatial extension.

A 11 $\mathsf{A\text{-}Star}(a, p, s) \wedge \mathsf{FAtt}(a) \rightarrow \exists a'[\,\mathsf{CAtt}(a') \wedge \mathsf{CP}(a', a) \wedge \mathsf{Star}(a', p, s)\,]$

If two spatial regions with the same coverage are spatially connected, then their spatial sum also denotes a spatial region with the same coverage.

A 12
$$\mathsf{Star}(a, p_1, s) \wedge \mathsf{Star}(a, p_2, s) \wedge \mathsf{CAtt}(a) \wedge \mathsf{C}(p_1, p_2) \rightarrow \exists p'[\mathsf{Star}(a, p', s) \wedge p' = \mathsf{external\text{-}boundary}(\mathsf{sum}(p_1, p_2))]$$

Given a spatial region r, every sub-region r' of r is also a region with the same coverage of r.

A 13 $\mathsf{Star}(a, p, s) \wedge \mathsf{CAtt}(a) \rightarrow \forall p'[\,\mathsf{P}(p', p) \rightarrow \mathsf{Star}(a, p', s)\,]$

Some attributes may be regarded as non-intersectable. It means that two spatial regions (or geographical features, or any combination of them) cannot be overlapped at a certain time instant if they are associated with non-intersectable attributes. For instance, one can say that a *arid* region and a *forest* feature cannot share the same spatial extension at the same time. These relationships are explicitly asserted as facts in the knowledge base, excepting for relating any combination of homogeneous coverage attributes and simple features, which are naturally non-intersectable (this is assured by the axioms A11 and A14). This relation is defined as follows.

D 4 $\mathsf{No\text{-}Intersection}(a_1, a_2) \equiv_{def} \mathsf{Att}(a_1) \wedge \mathsf{Att}(a_2) \wedge a_1 \neq a_2 \wedge \forall p_1 p_2 s[\,\mathsf{Instant}(s) \wedge \mathsf{Star}(a_1, p_1, s) \wedge \mathsf{Star}(a_2, p_2, s) \rightarrow \neg O(p_1, p_2)\,]$

A 14 $\mathsf{CAtt\text{-}Hom}(a_1) \wedge \mathsf{CAtt\text{-}Hom}(a_2) \wedge (a_1 \neq a_2) \rightarrow \mathsf{No\text{-}Intersection}(a_1, a_2)$

We have stated that when a *STAR* element is applied to a time interval it may represent the complete change of coverage in a spatial region or the entire creation of a geographical feature during a time interval. However, it may imply changes in other regions and features existing just before such interval and with which the new region cannot be intersected. For example, consider that the fact $\mathsf{No\text{-}Intersection}(forested, urbanised)$ is asserted in the knowledge base. In addition, suppose a dataset consisting of $\mathsf{Star}(a, p, i)$ elements representing urbanised regions created over regular time intervals (e.g. a week). Then, when some of these regions intersects a region which was forested at t_1 (where $t_1 = begin(i) - 1$), a new snapshot of this forested region at t_2 (where $t_2 = end(i)$) can be deduced if it is not stored explicitly in the database (assuming that the forested region at t_1 is also given). This is assured by the axioms A15 and A16.

A 15 $\mathsf{A\text{-}Star}(a, p, s) \wedge \mathsf{Interval}(s) \rightarrow \mathsf{Star}(a, p, \mathsf{e}(s)) \wedge \neg\exists p' t_1[\,(t_1 = \mathsf{b}(s) - 1) \wedge P(p', p) \wedge \mathsf{Star}(a, p', t_1)\,] \wedge \neg\exists a' p' t_2[\,(t_2 = \mathsf{e}(s)) \wedge \mathsf{O}(p, p') \wedge \mathsf{No\text{-}Intersection}(a, a') \wedge \mathsf{Star}(a', p', t_2)]$

A 16 $\mathsf{Star}(a, p, t_1) \wedge \mathsf{Instant}(t_1) \wedge \mathsf{Interval}(i) \wedge (t_1 = \mathsf{b}(i) - 1) \wedge \neg\exists a' p'[\,\mathsf{No\text{-}Intersection}(a, a') \wedge \mathsf{O}(p, p') \wedge \mathsf{Star}(a', p', i)\,] \rightarrow \mathsf{Star}(a, p, t_2) \wedge \mathsf{e}(i)$

We have shown different forms of data representation and the relationship between them, so that it is possible to deduce implicit data in reduced datasets. Although these implicit data can be deduced in the logical level (at reasoning time), a more efficient approach is to generate and store part of these explicitly, so that it can accessed more quickly as asserted facts at reasoning time. To define which data should be stored, we should consider both the processing time of generation and the storage space needed. In our implementation, depending on the type of data available to be used as input for the system, we submit them to a mechanism which enrich the original dataset by storing the following deduced data explicitly: maximal well-connected regions of some particular coverage, representing geographical features; and regions which has been changed as a consequence of the creation of other non-intersectable regions over a interval.

5 Logical Framework

We now present a logical framework we have named *RGP*, which is an acronym for *Reasoning about Geographical Processes*, comprising formal descriptions of space, time, events, processes, geographical objects (features) and their related aspects. This Section describes the basic syntax and semantics of the logical language employed in the framework. This language has been named $\Re$. Relevant predicates and logical relations employed in this framework shall be introduced in Sections 6 and 7.

The current version of this framework is restricted to dealing with processes which affects *homogeneous coverage regions* and *simple feature*. Future works shall include semantics to deal with different types of geographical data which is already supported by the data model described in Section 4. In this semantic model, we use the elements of type *feature type* and *regions coverage type* to represent possible types of region coverage attributes and feature attributes existing in the data model.

Events and processes are structured in terms of *types* and *tokens*. Event-types denote a certain kind of change which may affect a certain type of geographical feature. On the other hand, event-tokens denote particular occurrences of event types. These tokens are therefore associated with a specific time interval and a specific instance of a feature. For example, we may specify an event-type 'forest expansion' and an event-token to denote 'the expansion of the Amazon forest occurred between 01/01/2001 and 31/12/2001'. Similarly, process-types denotes a series of changes which take place in a certain feature-type, whilst a process-token denotes a particular instance of this type of process. Therefore tokens are said to *proceed* during a specific time interval and in a specific instance of a feature.

5.1 Syntax

The logical language $\Re$ used in the framework comprises variables of 10 *nominal types* which can be quantified over. The vocabulary of $\Re$ can be specified by a tuple $\mathcal{V} = \langle \mathcal{T}, \mathcal{I}, \mathcal{R}, \mathcal{F}, \mathcal{O}, \mathcal{L}, \mathcal{E}, \varSigma, \mathcal{P}, \varGamma \rangle$. These types and the variables used to assign them are listed below:

- Time Instants, $\mathcal{T} = \{..., t_i, ...\}$
- Time Intervals, $\mathcal{I} = \{..., i_i, ...\}$
- Spatial Regions, $\mathcal{R} = \{\varnothing, ..., r_i, ...\}$
- Geographical Features, $\mathcal{F} = \{..., f_i, ...\}$
- Homogeneous Coverage Types, $\mathcal{O} = \{..., o_i, ...\}$
- Simple Feature Types, $\mathcal{L} = \{..., l_i, ...\}$
- Event-types, $\mathcal{E} = \{..., e_i, ...\}$
- Event-tokens, $\varSigma = \{..., \sigma_i, ...\}$
- Process-types, $\mathcal{P} = \{..., \rho_i, ...\}$
- Process-tokens, $\varGamma = \{..., \gamma_i, ...\}$

The following *logical functions* are used to transfer information between distinct semantic types:

- $\mathsf{b}(i)$ and $\mathsf{e}(i)$ return, respectively, the time instant corresponding to the beginning and the end of the interval i.
- $\mathsf{dur}(\sigma)$ gives the duration of the interval of occurrence of the event-token σ.

- $\mathsf{ext}(f)$ returns a spatial region with the same spatial extension of a feature f.
- $\mathsf{f\text{-}type}(f)$ gives the type of the specified feature.
- $\mathsf{c\text{-}type}(r)$ returns the coverage type of the specified spatial region.

Several *predicates* and *logical relations* are also employed. They are described below:

- $t_1 < t_2$, $t_1 = t_2$, $t_1 \leq t_2$ are true, respectively, just in case the time term t_1 denotes a time earlier than the time denoted by t_2; t_1 denotes a time equal to t_2; t_1 denotes a time equal to or earlier than t_2.
- The following RCC relations between spatial regions: *connected* $\mathsf{C}(\alpha, \beta)$, *disconnected* $\mathsf{DC}(\alpha, \beta)$, *overlaps* $\mathsf{O}(\alpha, \beta)$, *externally connected* $\mathsf{EC}(\alpha, \beta)$, *part of* $\mathsf{P}(\alpha, \beta)$, *proper part of* $\mathsf{PP}(\alpha, \beta)$ and *equals to* $\mathsf{EQ}(\alpha, \beta)$, where α and β are region terms which may be either a region variable r_i, a term of the form $\mathsf{ext}(f_i)$ or the empty region constant $\varnothing$.
- $\mathsf{Holds\text{-}At}(\varphi, t)$ relation asserts that formula φ is true at the time instant denoted by t.
- $\mathsf{Occurs}(\sigma, i)$ relation means that the event-token σ occurs over the time interval denoted by i.
- $r_1 =_c r_2$, $r_1 \neq_c r_2$ are true, respectively, just in case the spatial region term r_1 denotes a region with the same type of coverage of region denoted by r_2; r_1 denotes a region with different type of coverage of r_2.

We also define the operator $f_1 \doteqdot f_2$, which is true if f_1 and f_2 are geographical features which have the same identity criteria. The identity criteria for a feature is defined in terms of the connectivity of its spatial extension over a time interval.

D 5 $$f_1 \doteqdot f_2 \leftrightarrow \forall it[\,\mathsf{b}(i) < t \leq \mathsf{e}(i) \wedge \mathsf{Holds\text{-}At}(\mathsf{EQ}(\mathsf{ext}(f_1), r_1), t-1) \wedge \mathsf{Holds\text{-}At}(\mathsf{EQ}(\mathsf{ext}(f_2), r_2), t) \rightarrow \mathsf{C}(r_1, r_2) \wedge \neg\exists r_3[\,r_1 =_c r_2 =_c r_3 \wedge \mathsf{DC}(r_3, r_2) \wedge \mathsf{DC}(r_3, r_1)\,]\,]$$

The following auxiliary functions are also employed to perform spatial calculations.

- $\mathsf{sum}(D)$ returns a spatial region which corresponds to the spatial sum of a set D of spatial regions.
- $\mathsf{area}(r)$ returns a number representing the area of a region r.

If φ and ψ are propositions of $\Re$, then so are the following:

- $\neg\varphi$, $(\varphi \wedge \psi)$, $(\varphi \vee \psi)$, $(\varphi \rightarrow \psi)$, $\forall v[\varphi]$

Where v is a variable of one of the nominal types described earlier.

5.2 Semantics

An *attributed geographic model* (AGM) is a structure $\mathcal{M} = \langle \mathfrak{G}, \mathcal{V}, \mathcal{A} \rangle$, where:

- $\mathfrak{G} = \langle \mathbb{R}^2, \langle T, \preceq \rangle, A, \mathcal{D} \rangle$ is a formal model of a geographic dataset:
  - $\mathbb{R}^2$ is the real plane, which will represent the geographic surface,[4]
  - T is a set of time points,
  - $\preceq$ is a total linear order over T,
  - A is a set of geographic attributes,
  - $\mathcal{D} \subseteq A \times \mathsf{Poly}(\mathbb{R}^2) \times \mathsf{Int}(T)$ represents the geographic attribute data as a set of tuples of the form $\langle a, p, s \rangle$, which correspond to the fact that attribute a holds for polygon p over interval s.[5]
    Here, $\mathsf{Poly}(\mathbb{R}^2)$ is the set of well-connected polygons over $\mathbb{R}^2$, and $\mathsf{Int}(T) = \{\langle t_1, t_2 \rangle \mid t_1, t_2 \in T \wedge t_1 \preceq t_2\}$ is the set of all intervals over the time sequence $\langle T, \preceq \rangle$.
- $\mathcal{V} = \langle \mathcal{T}, \mathcal{I}, \mathcal{R}, \mathcal{F}, \mathcal{O}, \mathcal{L}, \mathcal{E}, \Sigma, \mathcal{P} \rangle$, specifies the vocabulary of our representation language. Each element of this tuple is the set of all symbols of a given type (as specified in the syntax section).
- $\mathcal{A} = \langle a_{\mathcal{T}}, a_{\mathcal{I}}, a_{\mathcal{R}}, a_{\mathcal{F}}, a_{\mathcal{O}}, a_{\mathcal{L}}, a_{\mathcal{E}}, a_{\Sigma}, a_{\mathcal{P}}, a_{\Gamma} \rangle$ is a tuple of assignment functions specifying the denotations of all symbols in the vocabulary as follows:
  - $a_{\mathcal{T}} : \mathcal{T} \to T$, maps time point variables to time points.
  - $a_{\mathcal{I}} : \mathcal{I} \to \mathsf{Int}(T)$, maps interval variables to intervals.
  - $a_{\mathcal{R}} : \mathcal{R} \to \mathsf{Reg\text{-}Closed}(\mathbb{R}^2)$, maps region variables to regular closed regions of the plane.
  - $a_{\mathcal{F}} : \mathcal{F} \to (T \to \mathsf{Poly}(\mathbb{R}))$, maps each feature symbol to a function from time points to polygons (giving the spatial extension of the feature at each time point).
  - $a_{\mathcal{O}} : \mathcal{O} \to (T \to \mathsf{Reg\text{-}Closed}(\mathbb{R}^2))$, maps cover attributes to functions from time points to regular closed regions of the plane. This gives the extension of the region having a given type of coverage at each time point. In general this will be a multi-piece region.
  - $a_{\mathcal{L}} : \mathcal{L} \to (T \to 2^{\mathsf{Poly}(\mathbb{R})})$ maps each feature type to a function from time points to sets of polygons.
  - $a_{\mathcal{E}} : (\mathcal{E} \times \mathcal{F}) \to 2^{S}$, maps each combination of an event type symbol and a feature symbol to a set of non-overlapping intervals. These are the intervals during which there is an occurrence of the event type involving that feature.
  - $a_{\Sigma} : \Sigma \to (\mathcal{E} \times \mathcal{F} \times \mathsf{Int}(T))$, maps each event token symbol to a triple consisting of an event type, a feature (the participant) and an interval (the interval over which this particular event token occurs).

[4] Clearly, one might want to use a different coordinate system or a 2.5D surface model. For simplicity we just assume that the space is modelled by $\mathbb{R}^2$ but this could easily be changed without much modification to the rest of the semantics. It would affect the way that topological and metric properties are computed from the data, but our formal does not specify such implementation details.

[5] An interval may be *punctual*, if its beginning is the same as its end. Such intervals correspond to a single time points.

- $a_{\mathcal{P}} : (\mathcal{P} \times \mathcal{F}) \rightarrow 2^S$, maps each combination of a process type symbol and a feature symbol to a set of non-overlapping intervals. These are the intervals during which a process of the given type involving that feature *proceeds.*
- $a_{\Gamma} : \Gamma \rightarrow (\mathcal{P} \times \mathcal{F} \times \mathsf{Int}(T))$, maps each process token symbol to a triple consisting of an process type, a feature (the participant) and an interval (the interval over which this particular process token proceeds).

6 Geographical Process Definition

6.1 Events

Event types roughly correspond to natural language verbs. We use the sortal predicate $\mathsf{Event\text{-}Type}(e)$ to identify e as denoting an event type. The relation between an event-token (σ), its type (e) and its participant geographic feature (f) is represented by the predicate $\mathsf{Event}(\sigma, e, f)$. For convenience we also define:

D 6 $\mathsf{Participant\text{-}E}(\sigma, f) \equiv_{def} \exists e[\,\mathsf{Event}(\sigma, e, f)\,]$

D 7 $\mathsf{Event\text{-}Is\text{-}Of\text{-}Type}(\sigma, e) \equiv_{def} \exists f[\,\mathsf{Event}(\sigma, e, f)\,]$

We distinguish events which are purely spatial from others which are conceived as a geographical event. For example, the *shrinkage* of geographical features in general is considered a purely spatial event. However, when such shrinkage is associated with a specific type of feature it may be conceived as a geographical event. For example, when a feature *forest* shrinks, it may be interpreted as a geographical event, that is, a *deforestation* event. Therefore we define Event-types in terms of possible sub-types, i.e. *spatial change events* and *geographical events*:

D 8 $\mathsf{Event\text{-}Type}(e) \equiv_{def} \mathsf{Sp\text{-}Change\text{-}Event}(e) \vee \mathsf{Geo\text{-}Event}(e)$

We now define a relation which associates a spatial change event e_c, a geographical event e_g and a feature-type l. It means that when e_c occurs in a feature of type l it denotes the occurrence of a geographical event e_g.

D 9 $$\begin{aligned}\mathsf{Event\text{-}Feature}(e_c, l, e_g) \equiv_{def}\ & \mathsf{Sp\text{-}Change\text{-}Event}(e_c) \wedge \mathsf{Geo\text{-}Event}(e_g) \wedge \\ & (\mathsf{Occurs}(\sigma_1, i) \wedge \mathsf{Event}(\sigma_1, e_c, f) \wedge l = \mathsf{f\text{-}type}(f) \\ & \rightarrow \mathsf{Occurs}(\sigma_2, i) \wedge \mathsf{Event}(\sigma_2, e_g, f))\end{aligned}$$

The relation defined below is employed to associate two events which are interpreted as opposite each other, for example a *forestation event* and *deforestation event.* This relation is *symmetric* and *non-reflexive.*

D 10 $$\begin{aligned}\mathsf{Reverse\text{-}E}(e_1, e_2) \rightarrow\ & (\mathsf{Sp\text{-}Change\text{-}Event}(e_1) \wedge \mathsf{Sp\text{-}Change\text{-}Event}(e_2)) \\ & \vee \quad (\mathsf{Geo\text{-}Event}(e_1) \wedge \mathsf{Geo\text{-}Event}(e_2))\end{aligned}$$

For readability, we use this relation as a logical function $e_2 = \mathsf{reverse\text{-}e}(e_1)$.

Then we can now specify an axiom to ensure the integrity of the relation $\mathsf{Event\text{-}Feature}(e_c, l, e_g)$ when applied to reverse events.

A 17 $\mathsf{Event\text{-}Feature}(e_c, l, e_g) \leftrightarrow \mathsf{Event\text{-}Feature}(\mathsf{reverse\text{-}e}(e_c), l, \mathsf{reverse\text{-}e}(e_g))$

6.2 Processes

We model a process as a 'chunking' of events of the same type involving the same participant. We define some predicates and relation for processes which are similar to those defined for events, as follows. Process entities are distinguished by satisfying the predicate $\mathsf{Process\text{-}Type}(\rho)$. A process-token ($\gamma$) is related to its type (ρ) and participant (f), by the predicate $\mathsf{Process}(\gamma, \rho, f)$. We also define:

D 11 $\mathsf{Participant\text{-}P}(\gamma, f) \equiv_{def} \exists\gamma[\,\mathsf{Process}(\gamma, \rho, f)\,]$

D 12 $\mathsf{Process\text{-}Is\text{-}Of\text{-}Type}(\gamma, \rho) \equiv_{def} \exists f[\,\mathsf{Process}(\gamma, \rho, f)\,]$

We define the concept of *reverse process* which is analogous to the concept of reverse events. The *symmetric* and *non-reflexive* relation $\mathsf{Reverse\text{-}P}(\rho_1, \rho_2)$ is applied to associate two opposite process types. As defined for events, for readability, we also use this relation as a *logical function* $\rho_2 = \mathsf{reverse\text{-}p}(\rho_1)$.

The relation *Event Moves Process Forwards* $\mathsf{EMPF}(e, \rho)$ associates an event-type with a process-type. Asserting a fact using this relation means that if a process of type ρ is characterised by occurrences of events of type e, the process moves forwards. Similarly, the relation *Event Moves Process Backwards* $\mathsf{EMPB}(e, \rho)$ means that a process of type ρ moves backwards if it is characterised by occurrences of events of type e.

The axioms A18, A19 and A20 assure that the relations $\mathsf{EMPF}(e, \rho)$ and $\mathsf{EMPB}(e, \rho)$ are valid for reverse events and process.

A 18 $\mathsf{EMPF}(e, \rho) \leftrightarrow \mathsf{EMPB}(\mathsf{Reverse\text{-}E}(e), \rho)$

A 19 $\mathsf{EMPF}(e, \rho) \leftrightarrow \mathsf{EMPB}(e, \mathsf{Reverse\text{-}P}(\rho))$

A 20 $\mathsf{EMPF}(e, \rho) \leftrightarrow \mathsf{EMPF}(\mathsf{Reverse\text{-}E}(e), \mathsf{Reverse\text{-}P}(\rho))$

We can now present our approach to define when a process *proceeds*.

D 13
$$\begin{aligned}\mathsf{Proceeds}(\gamma, i) \equiv_{def}\ & \mathsf{Process}(\gamma, \rho, f) \wedge \\ & \forall t[\,\mathsf{b}(i) < t \leq \mathsf{e}(i) \rightarrow \exists i'e[\,t-1 = \mathsf{b}(i') \wedge t = \mathsf{e}(i') \wedge \\ & \qquad \mathsf{Occurs}(\mathsf{Event}(e, f), i') \wedge \mathsf{EMPF}(e, \rho)\,]\,]\end{aligned}$$

Similarly, we may also define the *proceeds* predicate in terms of the *reverse process* and the $\mathsf{EMPB}(e, \rho)$ relation, as follows.

D 14
$$\begin{aligned}\mathsf{Proceeds}(\gamma, i) \equiv_{def}\ & \mathsf{Process}(\gamma, \mathsf{reverse\text{-}p}(\rho), f) \wedge \\ & \forall t[\,\mathsf{b}(i) < t \leq \mathsf{e}(i) \rightarrow \exists i'e[\,t-1 = \mathsf{b}(i') \wedge t = \mathsf{e}(i') \wedge \\ & \qquad \mathsf{Occurs}(\mathsf{Event}(e, f), i') \wedge \mathsf{EMPB}(e, \rho)\,]\,]\end{aligned}$$

The relation *Process Component* is true in case an event σ is said to be a component of a process γ. This is defined as follows.

D 15
$$\begin{aligned}\mathsf{Proc\text{-}Comp}(\sigma, \gamma) \equiv_{def}\ & \exists ii'[\mathsf{Proceeds}(\gamma, i) \wedge \mathsf{Occurs}(\sigma, i') \wedge \\ & \mathsf{b}(i) \leq \mathsf{b}(i') \leq \mathsf{e}(i') \leq \mathsf{e}(i) \wedge \\ & \mathsf{Participant\text{-}E}(\sigma, f) \wedge \mathsf{Participant\text{-}P}(\gamma, f)\end{aligned}$$

7 Defining Properties of Processes

In this section we present approaches to represent properties of processes. We describe a set of properties which may be applied to a variety of geographical processes: Initiation, Cessation, Acceleration, Deceleration and Proceeding Constantly.

7.1 Process Initiation and Cessation

Using the predicate $\mathsf{Proceeds}(\gamma, i)$, which delimits temporal boundaries of a process instance, the definition of process initiation and cessation is straightforward:

D 16 $\mathsf{Initiation}(\gamma, t) \equiv_{def} \exists i[\mathsf{Proceeds}(\gamma, i) \wedge t = \mathsf{b}(i)]$

D 17 $\mathsf{Cessation}(\gamma, t) \equiv_{def} \exists i[\mathsf{Proceeds}(\gamma, i) \wedge t = \mathsf{e}(i)]$

7.2 Process Acceleration, Deceleration and Constant Proceeding

A process acceleration, deceleration and constant proceeding may be defined in many forms, which depend on the type of geographical process being investigated and which spatial process is associated to their activity. We now present an approach to define these properties applied to geographical processes whose activities are based on the expansion and shrinkage of geographical features.

Spatial Expansion and Shrinkage events are defined in terms of mereological relationships between the spatial extension of a feature before and after the occurrence of such event. This is as follows.

D 18 $\mathsf{Occurs}(\mathsf{Event}(\mathit{expansion}, f), i) \equiv_{def}$

$$\begin{aligned}&\mathsf{Holds\text{-}At}(\mathsf{EQ}(\mathsf{ext}(f), r_b), \mathsf{b}(i) - 1) \quad \wedge\\ &\mathsf{Holds\text{-}At}(\mathsf{EQ}(\mathsf{ext}(f), r_e), \mathsf{e}(i)) \quad \wedge\\ &(\mathsf{PP}(r_b, r_e) \vee \mathsf{PO}(r_b, r_e) \vee \mathsf{EC}(r_b, r_e))\end{aligned}$$

D 19 $\mathsf{Occurs}(\mathsf{Event}(\mathit{shrinkage}, f), i) \equiv_{def}$

$$\begin{aligned}&\mathsf{Holds\text{-}At}(\mathsf{EQ}(\mathsf{ext}(f), r_b), \mathsf{b}(i) - 1) \quad \wedge\\ &\mathsf{Holds\text{-}At}(\mathsf{EQ}(\mathsf{ext}(f), r_e), \mathsf{e}(i)) \quad \wedge\\ &(\mathsf{PP}(r_e, r_b) \vee \mathsf{PO}(r_b, r_e) \vee \mathsf{EC}(r_b, r_e))\end{aligned}$$

Expansion and Shrinkage Rates should be defined in order to define process properties of acceleration, deceleration and constant proceeding. These rates could be defined in several ways, such as in terms of measurement of absolute area or as a percentage. We now present an approach to calculate these rates. Then it shall be employed to illustrate how we can define these process properties. We specify a generic relation $\mathsf{Rate}(\sigma, x)$, which assigns to x the rate calculated for a spatial change event σ.

Expansion Rate

D 20 $\mathsf{Rate}(\sigma, x) \leftarrow$

$$\begin{array}{l}
\mathsf{Occurs}(\mathsf{Event}(expansion, f), i) \quad \wedge \\
\mathsf{Holds\text{-}At}(\mathsf{EQ}(\mathsf{ext}(f), r_b), \mathsf{b}(i) - 1) \quad \wedge \\
\mathsf{Holds\text{-}At}(\mathsf{EQ}(\mathsf{ext}(f), r_e), \mathsf{e}(i)) \quad \wedge \\
D = \{\, r_1, r_2, ..., r_n \mid \neg\mathsf{P}(r_i, r_b) \wedge \mathsf{P}(r_i, r_e)\} \quad \wedge \\
x = (\mathsf{area}(\mathsf{sum}(D))/\mathsf{area}(r_b))
\end{array}$$

Shrinkage Rate

D 21 $\mathsf{Rate}(\sigma, x) \leftarrow$

$$\begin{array}{l}
\mathsf{Occurs}(\mathsf{Event}(shrinkage, f), i) \quad \wedge \\
\mathsf{Holds\text{-}At}(\mathsf{EQ}(\mathsf{ext}(f), r_b), \mathsf{b}(i) - 1) \quad \wedge \\
\mathsf{Holds\text{-}At}(\mathsf{EQ}(\mathsf{ext}(f), r_e), \mathsf{e}(i)) \quad \wedge \\
D = \{\, r_1, r_2, ..., r_n \mid \mathsf{P}(r_i, r_b) \wedge \neg\mathsf{P}(r_i, r_e)\} \quad \wedge \\
x = (\mathsf{area}(\mathsf{sum}(D))/\mathsf{area}(r_b))
\end{array}$$

Process Acceleration, Deceleration and Constant Proceeding can now be defined by using the relation we provided to calculate the changing rate for the required spatial change events. These properties can be applied to any geographical process which is composed by geographical events denoted by the occurrence of such spatial change events. These properties are defined as follows.

D 22 $\mathsf{acceleration}(\gamma, i) \equiv_{def} \mathsf{Proceeds}(\gamma, i) \quad \wedge$

$$\begin{array}{l}
\mathsf{Proc\text{-}Comp}(\sigma_1, \gamma) \wedge \mathsf{Proc\text{-}Comp}(\sigma_2, \gamma) \quad \wedge \\
\mathsf{e}(\mathsf{dur}(\sigma_1)) < \mathsf{e}(\mathsf{dur}(\sigma_2)) \quad \wedge \\
\mathsf{Rate}(\sigma_1, x_1) \wedge \mathsf{Rate}(\sigma_2, x_2) \rightarrow x_2 > x_1
\end{array}$$

D 23 $\mathsf{deceleration}(\gamma, i) \equiv_{def} \mathsf{Proceeds}(\gamma, i) \quad \wedge$

$$\begin{array}{l}
\mathsf{Proc\text{-}Comp}(\sigma_1, \gamma) \wedge \mathsf{Proc\text{-}Comp}(\sigma_2, \gamma) \quad \wedge \\
\mathsf{e}(\mathsf{dur}(\sigma_1)) < \mathsf{e}(\mathsf{dur}(\sigma_2)) \quad \wedge \\
\mathsf{Rate}(\sigma_1, x_1) \wedge \mathsf{Rate}(\sigma_2, x_2) \rightarrow x_2 < x_1
\end{array}$$

D 24 $\mathsf{constant}(\gamma, i) \equiv_{def} \mathsf{Proceeds}(\gamma, i) \quad \wedge$

$$\begin{array}{l}
\mathsf{Proc\text{-}Comp}(\sigma_1, \gamma) \wedge \mathsf{Proc\text{-}Comp}(\sigma_2, \gamma) \quad \wedge \\
\mathsf{e}(\mathsf{dur}(\sigma_1)) < \mathsf{e}(\mathsf{dur}(\sigma_2)) \quad \wedge \\
\mathsf{Rate}(\sigma_1, x_1) \wedge \mathsf{Rate}(\sigma_2, x_2) \rightarrow x_1 = x_2
\end{array}$$

8 Conclusions and Further Work

We have presented a representational model and a reasoning mechanism to analyse evolving geographical features and their relationship to geographical processes, in order to identify manifestations of certain properties which may be ascribed to these processes. We also described an approach to modelling the spatio-temporal data upon which the logical framework is grounded. This approach provides flexibility for storing datasets distributed in a variety of formats.

Further investigations shall be conducted in order to add the capability of storing spatial elements of several kinds of geometrical types, instead of restricting to polygonal types.

We introduce a set of properties which can be associated with several geographical processes, however future works shall include additional properties which should also be applicable to a variety of processes. The current version of the logical framework presented in this paper is restricted to dealing with processes which affects homogeneous coverage regions and simple features. Therefore, further works shall enrich its semantics to deal with different types of geographical data which is already supported by the proposed data model.

Of particular interest for further works is the incorporation of approaches to handling vagueness in the proposed reasoning mechanism. Geographical processes may be affected by vagueness in many ways, specially to defining spatial and temporal boundaries tanking into account different possible interpretations. This includes issues related to spacial and temporal aggregation and the treatment of information granularity.

References

1. Batty, M.: Geocomputation Using Cellular Automata. Geocomputation, 95–126 (2000)
2. Clarke, C.K., Brass, A.J., Riggan, J.P.: A Cellular Automaton Model of Wildfire Propagation and Extinction. Photogrammetric Engineering and Remote Sensing 60(11), 1355–1367 (1994)
3. Claramunt, C., Parent, C., Thériault, M.: Design patterns for spatiotemporal processes. Searching for Semantics: Data Mining, Reverse Engineering, 415–428 (1997)
4. Claramunt, C., Theriault, M.: Toward semantics for modelling spatio-temporal processes within GIS. In: Advances in GIS Research I, pp. 27–43 (1996)
5. Claramunt, C., Thriault, M., Parent, C.: A qualitative representation of evolving spatial entities in two-dimensional spaces. In: Innovations in GIS V, pp. 119–129 (1997)
6. Cohn, A.G., Bennett, B., Gooday, J., Gotts, N.: RCC: a calculus for region-based qualitative spatial reasoning. GeoInformatica 1, 275–316 (1997)
7. Crooks, A.: Exploring cities using agent-based models and GIS. In: Proceedings of the Agent 2006 Conference on Social Agents: Results and Prospects, Citeseer (2006)
8. Devaraju, A., Kuhn, W.: A Process-Centric ontological approach for integrating Geo-Sensor data. In: 6th International Conference on Formal Ontology in Information Systems, FOIS 2010 (2010)
9. Erol, K., Levy, R., Wentworth, J.: Application of agent technology to traffic simulation. In: Complex Systems, Intelligent Systems and Interfaces, Nimes, France (May 1998)
10. Frank, A.U., Campari, I., Formentini, U. (eds.): GIS 1992. LNCS, vol. 639. Springer, Heidelberg (1992)
11. Galton, A.: Desiderata for a spatio-temporal geo-ontology. Spatial Information Theory, 1–12 (2003)
12. Galton, A.: Experience and history: Processes and their relation to events. Journal of Logic and Computation 18(3), 323–340 (2007)

13. Galton, A.: A formal theory of objects and fields. In: Montello, D.R. (ed.) COSIT 2001. LNCS, vol. 2205, pp. 458–473. Springer, Heidelberg (2001)
14. Galton, A.: Spatial and temporal knowledge representation. Earth Science Informatics 2(3), 169–187 (2009)
15. Grenon, P., Smith, B.: SNAP and SPAN: towards dynamic spatial ontology. Spatial Cognition & Computation, 69–104 (2004)
16. Hornsby, K., Egenhofer, M.J.: Identity-based change: A foundation for spatio-temporal knowledge representation. International Journal of Geographical Information Science 14, 207–224 (2000)
17. Ohgai, A., Gohnai, Y., Watanabe, K.: Cellular automata modeling of fire spread in built-up areas–A tool to aid community-based planning for disaster mitigation. Computers, Environment and Urban Systems 31(4), 441–460 (2007)
18. Parker, D.C., Manson, S.M., Janssen, M.A., Hoffmann, M.J., Deadman, P.: Multi-agent systems for the simulation of land-use and land-cover change: A review. Annals of the Association of American Geographers 93(2), 314–337 (2003)
19. Randell, D.A., Cui, Z., Cohn, A.G.: A spatial logic based on regions and connection. In: KR 1992, pp. 165–176 (1992)
20. Shimabukuro, Y., Duarte, V., Anderson, L., Valeriano, D., Arai, E., de Freitas, R., Rudorff, B., Moreira, M.: Near real time detection of deforestation in the Brazilian Amazon using MODIS imagery. Revista Ambiente & Água-An Interdisciplinary Journal of Applied Science 1(1) (2006)
21. Wainwright, J.: Can modelling enable us to understand the rôle of humans in landscape evolution?. Geoforum 39(2), 659–674 (2008)
22. Walter, V.: Object-based classification of remote sensing data for change detection. ISPRS Journal of Photogrammetry and Remote Sensing 58(3-4), 225–238 (2004)
23. White, R., Engelen, G.: Cellular automata and fractal urban form: a cellular modelling approach to the evolution of urban land-use patterns. Environment and Planning A 25(8), 1175–1199 (1993)
24. Wolter, F., Zakharyaschev, M.: Spatio-temporal representation and reasoning based on rcc-8. In: Proceedings of the Seventh Conference on Principles of Knowledge Representation and Reasoning, KR 2000, pp. 3–14. Morgan Kaufmann, San Francisco (2000)
25. Worboys, M., Hornsby, K.: From objects to events: GEM, the geospatial event model. In: Egenhofer, M.J., Freksa, C., Miller, H.J. (eds.) GIScience 2004. LNCS, vol. 3234, pp. 327–343. Springer, Heidelberg (2004)

DO-ROAM: Activity-Oriented Search and Navigation with OpenStreetMap

Mihai Codescu[1], Gregor Horsinka[1], Oliver Kutz[2],
Till Mossakowski[1,2], and Rafaela Rau[1]

[1] DFKI GmbH Bremen
[2] Research Center on Spatial Cognition (SFB/TR 8),
University of Bremen, Germany

Abstract. We develop a web service focusing on finding places not (only) by their address, but by systematically relating the places to activities that a person could perform there. This is helpful if a person wants to explore a new city, or plans leisure activities. OpenStreetMap provides a rich set of tags that can be used for activity-oriented search. We propose the use of several ontologies that are related to each other using matching tools to cope with the evolving nature of the tags available in social media.

1 Introduction

OpenStreetMap has evolved into a rich source of geodata that in some aspects (like e.g. the level of detail for certain pedestrian lanes) even gets ahead of Google maps. When searching and navigating through a map portal like `http://www.openstreetmap.org`, semantic metadata could greatly help with providing an *intention* and *activity*-based access to the data. In the case of OpenStreetMap, the metadata is provided in the form of *tags* that are entered into the database in a Social Web and wiki-like manner. Metadata obtained through such Social Web, collaborative and community based efforts have specific characteristics, namely evolve in a bottom-up way, contain a lot of noise (typos, redundancies, etc.) and are subject to constant change. A main challenge now is how to use such metadata in flux in a meaningful way for an activity-based search and navigation tool. In this paper, we use ontologies and (semi-automatically generated) ontology mappings for bridging the gap between (a single) user's intentions and the (community generated) metadata tags. This approach provides a relatively simple, yet effective solution to the generally rather hard problem of how to relate data to ontologies (see [15]).

Based on this, we have developed an open source tool—DO-ROAM[1]—which is a prototype providing, beyond the usual search facilities inherited from the OpenStreetMap portal, an ontology-based search for *located activities* and *opening hours*.

[1] Freely available at `www.do-roam.org`

C. Claramunt, S. Levashkin, and M. Bertolotto (Eds.): GeoS 2011, LNCS 6631, pp. 88–107, 2011.

1.1 Related Work

The GeoShare project did pioneering work on ontology-based integration of geodata sources and services [9]. Their system performs ontological, spatial and temporal reasoning when processing user queries. However, as the authors note, "all application ontologies have in common that they are based on the same vocabulary". We here follow a more flexible approach based on ontology matching. Moreover, the authors of the GeoShare/Buster software told us that it is not used by any web service, and it would be a great effort to get the software running again. Indeed, we have the impression that the user interface was too complex—a stripped down version, however without the ontology-based search, is still online[2].

Google maps[3] obviously uses a mixture of full-text search and search in a taxonomy of categories; unfortunately, only parts of the taxonomy are openly available. While full-text search can in some cases provide extra value, sometimes it can also produce misleading results. For example, when entering "new york barber restaurant" or "new york barber near restaurant", you get results in the category "restaurant" for which the word "barber" occurs in a related text document (or vice versa), but only few restaurants near barbers. Searching for "charging station" in most cities does not deliver any results at all.

In comparison, whilst searching for "charging station" in OpenStreetMap does in fact not deliver any results at all, OSM's internal data is open and publicly accessible, and thus better search methods, employing e.g. the OSM tags, can be utilised in order to semantically enrich queries. Thus, we intend to realise an activity-oriented search where several activities can be combined, thereby leading to various possibilities for searching for nearby places, or for the restriction of a search to certain opening hours.

Google city tours[4] suggests touristic tours starting from a given point; however, the user cannot enter specific activities. Other works that stress the importance of activities and actions in GIS include [10,17], who argue that in order to make geographical information really useful, corresponding ontologies would have to be designed with a focus on human activities rather than being 'static and entity-based'. Moreover, similar to our approach mapping metadata tags to activities in an ontology, e.g. [10] proposes to exploit textual descriptions of activities in order to derive domain ontologies. However, unlike the present paper, these works do neither employ statistical matching methods to link these two layers, nor do they use the OWL language, nor apply ideas from recent progress in ontology-based data access.

Our work has been much inspired by an activity-oriented interactive route planning system [19], see Fig. 1 and `http://www.digitaltravelmate.net`. This system allows the user to specify interesting locations via holiday activities, and routes are planned along locations where the selected activities can be performed.

[2] `http://www.geoshare.umwelt.bremen.de`

[3] `http://maps.google.com`

[4] `http://citytours.googlelabs.com`

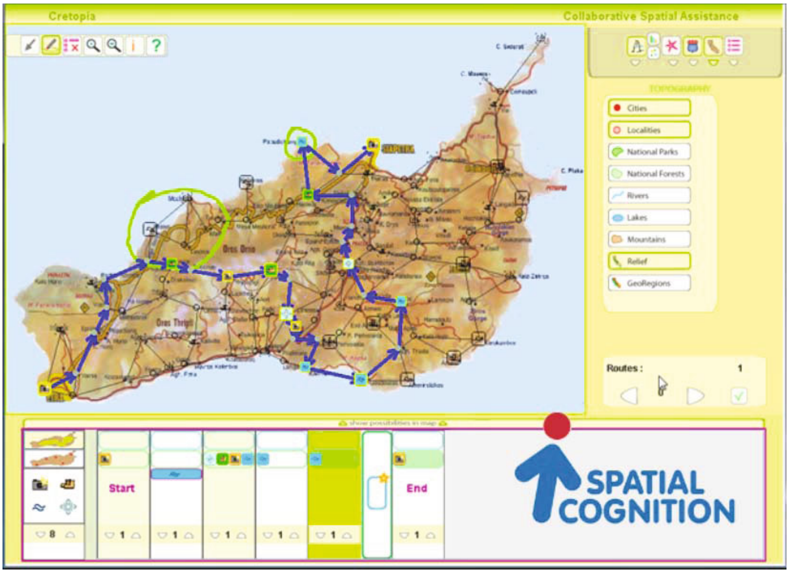


Fig. 1. Activity-oriented interactive route planning system

Routes can also be interactively corrected. However, the system is based on a *fixed* fictitious map, and on a small predefined set of activities.

1.2 Organisation of the Paper

The aim of our work reported in this paper is to provide activity-oriented map search and navigation with an *evolving* set of activities based on OpenStreetMap's tags, which are continuously changing due to the wiki-nature of OpenStreetMap.

The paper is organised as follows. Section 2 provides a set of use cases, and in Section 3 we describe the overall architecture of our DO-ROAM system. The main parts of this system are then described in further detail in the subsequent sections: Section 4 recalls some technical background on the web ontology language OWL, Section 5 introduces the ontology of activities and Section 6 the ontology of OpenStreetMap tags, Section 7 discusses the mapping between these two ontologies, and Section 8 finally contains a discussion on how the ontology search is integrated with the actual representation of data. Section 9 concludes and discusses future work.

2 Motivating Use Cases

In this section, we describe some scenarios in which people search for locations where certain activities take place and for routes that include such activities. One important such application scenario is electric mobility, in particular due to the limited battery reach of electric automobiles and relatively long charging times. Notice that we assume that the map presented to the user is taken from OpenStreetMap; the approach is, however, flexible and it could also use multiple data sources. [1] provide different navigation scenarios based on an ontology for GIS, which inspired the format of our use cases.

- *Scenario 1:* Alan spends some days in a city and he want to charge his electric car. First, he wants to know where he can find a charging station. Second, he is interested in which activities he could do within walking distance from the location of the charging station.
- *Scenario 2:* Betty is new in town. She wants to know which activities are offered within her neighbourhood. She also knows she will be getting hungry soon, so she searches for all restaurants close to home and which will be still open within the next two hours.
- *Scenario 3:* Maria wants to visit her friend. On her way she needs to stop at a supermarket, an ATM and a post office. She needs a system which will generate and present to her a route, including all these stops, in any order. She also wants to be able to modify the resulting route.
- *Scenario 4:* Tom wants to travel from A to B. He wants to take the most scenic route possible. He also wants to see displayed all places of his interest within a certain area that he can choose and modify. Furthermore, he wants to get a route suggestion which is still flexible and can be modified at a later stage.

3 General Tool Architecture of DO-ROAM

We have designed and implemented the prototype of a tool DO-ROAM for answering such requests and for assisting the users in spatio-temporal planning of activities. DO-ROAM is an acronym for *Data and Ontology driven Route-finding Of Activity-oriented Mobility.*

Currently, only the search component is implemented and the route finding integration is in progress. Therefore, only Scenarios 1 and 2 from those mentioned in Section 2 are currently supported. The general GUI of the tool is illustrated in Fig. 2. The tool displays a map (based on OpenStreetMap) with a zoom functionality which allows the user to focus on a certain area of interest. Searching for locations which allow to perform desired activities can be done either in a guided way, using the ontology navigation bar on the left of the map, or in a less contrained way, using a text field for introducing the query. In the latter case, address and opening hours can also be taken into account.

The tool is implemented as a Web application, using Ruby on Rails[5], a popular and powerful web application framework. We have built our tool on top of the existing Rails portal for OpenStreetMap[6]. It must solve a data integration problem in the sense that the way the OpenStreetMap data is represented

[5] http://rubyonrails.org/

[6] http://wiki.openstreetmap.org/wiki/The_Rails_Port

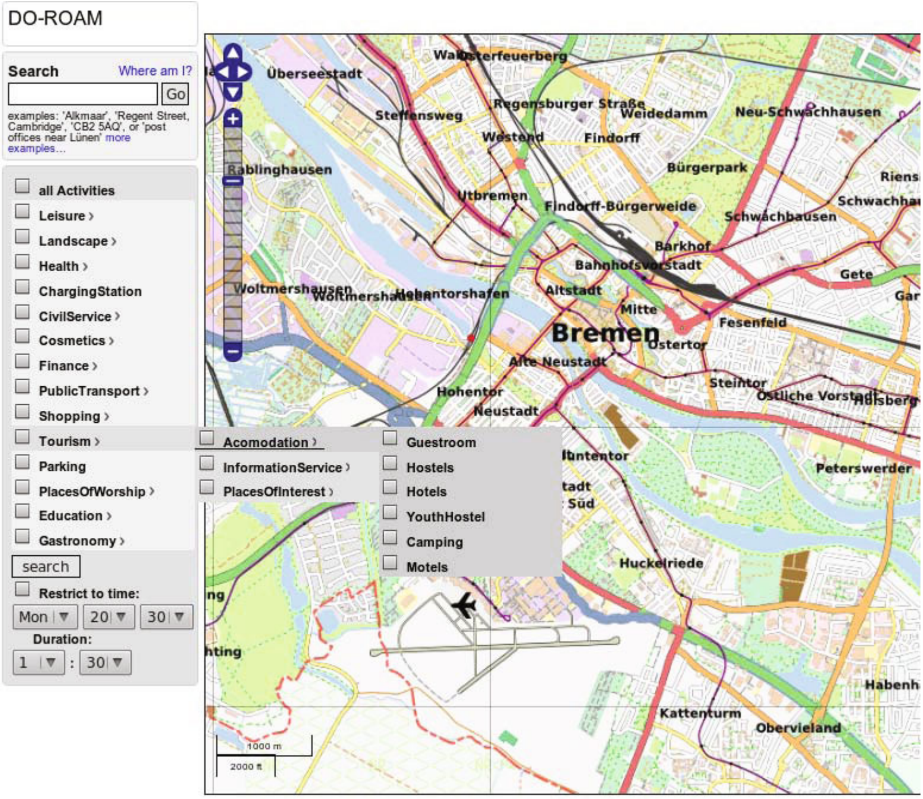


Fig. 2. User interface of DO-ROAM prototype

should not be directly visible to the user, and the interaction with the user should be as facile as possible. The solution employed is *ontology-based data access* (OBDA) [16,4], where the domain of interest is modelled as an ontology which is connected with the data in a way that allows queries expressed in terms of the ontology to be translated to queries in the database. Therefore, we introduced an ontology of spatially located activities playing a central role and connecting the user interface with the data integration management system. The ontology will be discussed in detail in Section 5. Interestingly, the access to data is achieved by introducing another ontology for OpenStreetMap tags, which will be presented in Section 6. The two ontologies are connected via an ontology mapping, which relates the concepts/roles in the ontology of activities with corresponding concepts/roles in the ontology of OSM tags. We will discuss some of the fundamentals of ontology mappings and the means for generating such mappings automatically in Section 7. Moreover, we give a brief intuition on OBDA and its implementation in our tool in Section 8. Finally, the results of the queries are displayed on a map using OpenStreetMap layers: each location of a certain activity is marked with a distinctive icon. This is realized dynamically in the sense that the markers are only introduced for locations within the current view.

The architecture of our approach is depicted in Fig. 3.

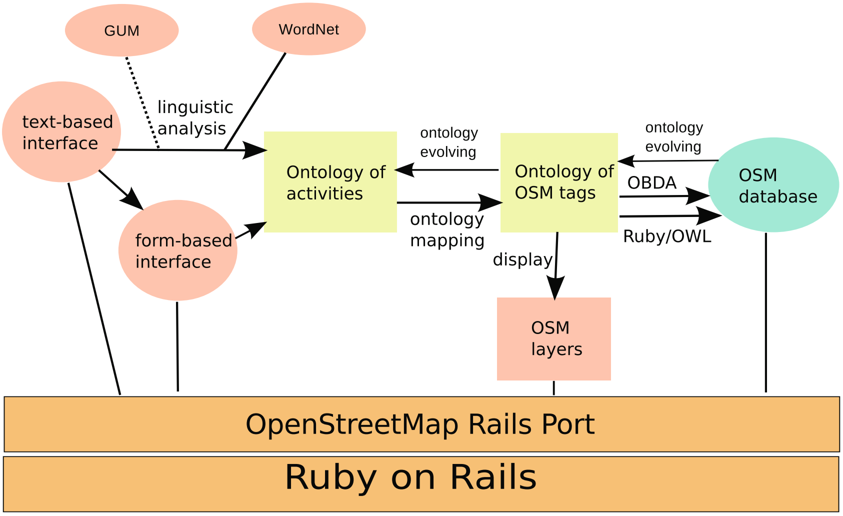


Fig. 3. Architecture of DO-ROAM's activity-based search

The user interaction is handled via two alternative interfaces, which we will now motivate and describe in some detail. The first one is a simple text-based interaction for free search, similar to the one existing in tools like Google Maps or OpenStreetMap, while the second provides a better structuring of the query with the help of a Web form. While the first interface seems more intuitive, the second has the advantage that it is easier to relate with the concepts in the ontology of activities. In the case of the former, the text input by the user needs to undergo a process of linguistic analysis which extracts from the query the concepts which are matched.

For the linguistic analysis we currently use WordNet [5] synsets. That is, the user need not exactly match the concept names of the activities ontology with his query, but can also enter synonyms. For example, take "eating place". If you enter "eating place New York" into Google maps, you only get a few restaurants, however, if you enter "restaurant New York", you get plenty of restaurants. With our connection to WordNet, we get the same set of restaurants for both queries since WordNet knows that "eating place" and "restaurant" are synonyms. Another example would be "clothing": when typing in "clothing London" into Google Maps, you get plenty of results; however, when entering "dress" or "vesture", Google Maps delivers very few outputs; in case of "vesture", not even one. WordNet in contrast knows all words as synonyms of clothing.

The linguistic analysis also needs to divide the user's input into class names (from the activities ontology), address parts, and time information, and it even needs to infer role names in some cases (like the case with restaurants having a cuisine of a given nationality). This is currently done with a simple matching in a comma-separated list, then the synonyms of the activity are obtained from WordNet and used to generate a list of queries. Addresses are resolved using the search engines of OSM, e.g. Nominatim[7]. In the future, in particular in connection with way finding, we will also use the linguistic ontology GUM [3], since it provides a more detailed semantics for linguistic spatial expressions.

We will now illustrate the way the user interacts with the tool with a stepwise description for the case of Scenario 1 in Section 2 . First, Alan finds the city by using the free search text field. Afterwards, he chooses "Charging Station" in the ontology navigation bar. He can then select one charging station of his choice and zoom in to the desired scale (which is indicated using the functionality of the OSM Rails Port in the bottom-left of the map), obtaining thus an estimate of the area reachable by foot. Then he can get displayed markers for the locations of the free time activities or even all possible activities using again the navigation bar. The results can be seen in Fig. 4. There is one charging station, marked with a plug, and various activities marked with different icons.

In Scenario 2, Betty searches for her address using the free search text field, then she selects in the ontology navigation bar "Gastronomy/Restaurants" and enters the current time and two hours for duration. The results of this query are shown in Fig. 5. There are several restaurants, a café (at the lower right corner) and some fastfoods.

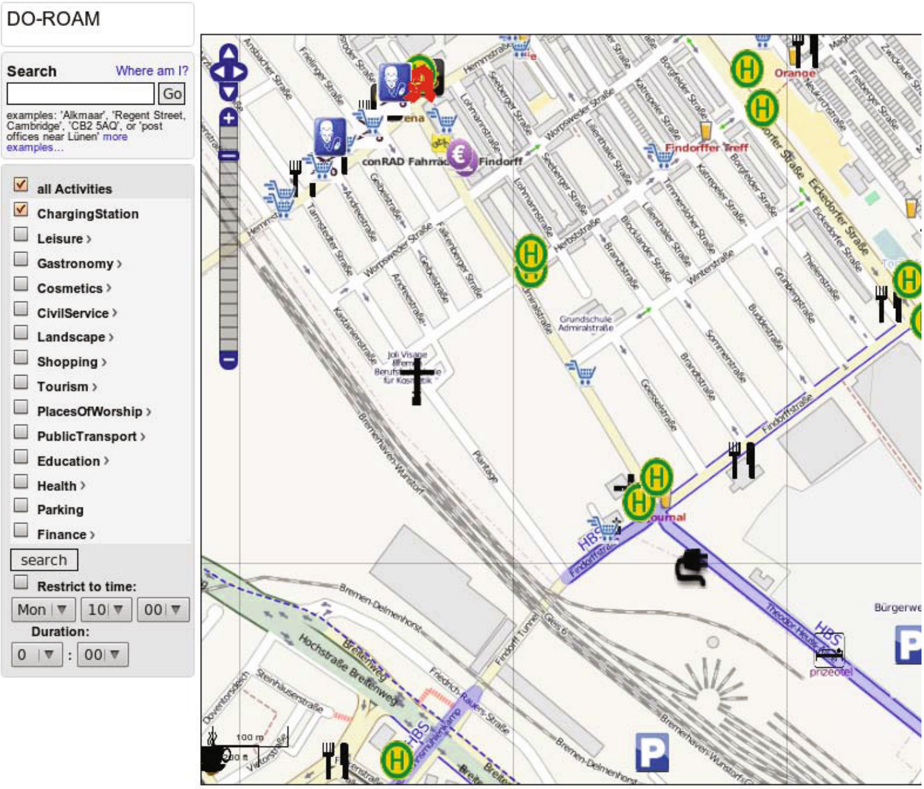
4 Ontologies and the OWL Language

Ontologies are formal descriptions of the concepts in a certain domain of discourse and can be informally understood as fixing a meaning for the terms of a particular field. Ontologies are used in artificial intelligence, the semantic web, systems engineering, software engineering, biomedical informatics, library science, enterprise bookmarking, and information architecture as a form of knowledge representation about the world or some part of it. Domain ontologies are typically formulated in the web ontology language OWL[8]. The relation of our domain ontology introduced in the next section to a suitable foundational ontology (typically formulated in a richer logical language) is left for future work.

Formally, an OWL ontology signature consists of sets of *atomic concepts*, *roles* and *individuals*, which fix the vocabulary. Sentences that can be expressed are of two types: TBox sentences are subsumption relations between concepts which are defined inductively from atomic concepts using the universal concept, the empty concept, unions, disjunctions, negations and universal and existential quantification over roles. ABox sentences contain assertions saying that certain individuals belong to certain complex concepts expressible in the vocabulary. Since the ontologies we use here do not contain individuals, we will concentrate on presenting TBox sentences.

[7] http://wiki.openstreetmap.org/wiki/Nominatim

[8] http://www.w3.org/TR/owl2-overview/



DO-ROAM

Search Where am I?

Go

examples: 'Alkmaar', 'Regent Street, Cambridge', 'CB2 5AQ', or 'post offices near Lünen' more examples...

☑ all Activities
☑ ChargingStation
☐ Leisure ›
☐ Gastronomy ›
☐ Cosmetics ›
☐ CivilService ›
☐ Landscape ›
☐ Shopping ›
☐ Tourism ›
☐ PlacesOfWorship ›
☐ PublicTransport ›
☐ Education ›
☐ Health ›
☐ Parking
☐ Finance ›

search

☐ Restrict to time:

Mon | 10 | 00

Duration:

0 : 00

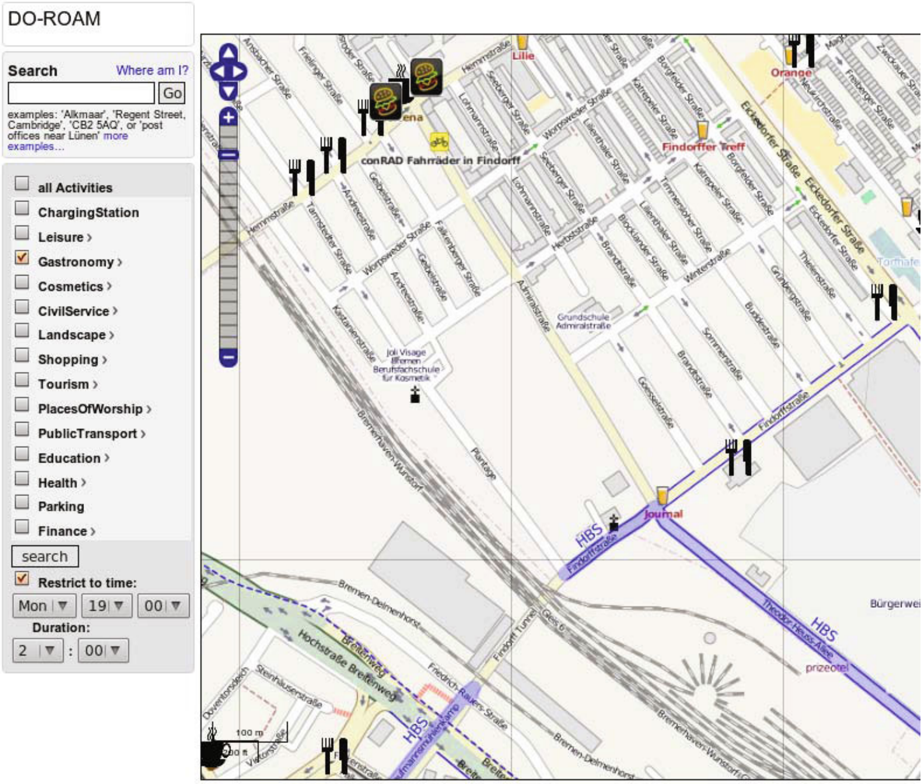
Fig. 4. User interface of DO-ROAM prototype: looking for activities near charging stations

Several syntaxes have been designed for ontology languages; in this paper we prefer to use Manchester OWL syntax [7] which provides, for the fragment corresponding to the description logic $\mathcal{ALC}$, the following grammar for concepts:

$$C ::= A \mid Thing \mid Nothing \mid C \ and \ C \mid C \ or \ C \mid not \ C \mid R \ some \ C \mid R \ all \ C$$

where R is a role and A is an atomic concept.

The semantics is set-theoretical: an interpretation I consists of a non-empty set W (the universe) and an interpretation function $.^I$ which assigns a subset of the universe to each atomic concept, a binary relation to each role and an element of the universe to each individual. The interpretation extends from atomic concepts to complex concepts in the expected set-theoretic way following the grammar, more precisely: the top concept $Thing$ is interpreted as the universe W, $Nothing$ as the empty set (bottom concept), a *conjunction* $C \ and \ D$ by the intersection of the interpretations for C and D, a *disjunction* $C \ or \ D$ by



☐ all Activities
☐ ChargingStation
☐ Leisure ›
☑ Gastronomy ›
☐ Cosmetics ›
☐ CivilService ›
☐ Landscape ›
☐ Shopping ›
☐ Tourism ›
☐ PlacesOfWorship ›
☐ PublicTransport ›
☐ Education ›
☐ Health ›
☐ Parking
☐ Finance ›

☑ Restrict to time: Mon ▼ 19 ▼ 00 ▼
Duration: 2 ▼ : 00 ▼

Fig. 5. User interface of DO-ROAM prototype: looking for restaurants open in the next two hours

the union of the interpretations, *not* C by set-theoretic complement, and finally *universal* (R *all* C) and *existential* (R *some* C) *role restrictions* as follows:

$$(R \textit{ all } C)^I = \{x \in W \ \mid\ \forall y \in W \,.\, R^I(x, y) \text{ implies } y \in C^I\}$$

and

$$(R \textit{ some } C)^I = \{x \in W \ \mid\ \exists y \in W \,.\, R^I(x, y) \text{ and } y \in C^I\}$$

Two ontologies can be related by an *ontology mapping*, sending atomic concepts, roles and individuals of the source ontology to (not necessarily atomic) concepts, roles and individuals of the target ontology. Among many other applications, ontology mappings are important for extracting modules from large ontologies.

5 An Ontology of Spatially Located Activities

Since the scenarios presented in Section 2 are centred on activities, we develop an ontology of *spatially located activities*, which means that the concepts of the ontology refer to locations where a certain activity takes place. This provides an abstraction level from the representation of the data in the databases and thus the user can express queries using a vocabulary closer to natural language. Notice that from an ontological perspective, the ontology developed is a *task ontology*: here the main motivation is not to create and specify a model of a domain, but to solve a well-defined task, namely, searching locations.

We found some interesting guidelines for building ontologies in [12]. In particular, they say

> We advise only creating hierarchies when necessary for describing the domain [. . .]. The modeller should consider whether an alternative relationship can be used instead. [12]

This means that instead of subclass relations, sometimes it is more useful to use roles. For example, instead of turning “French” into subclass of “Cuisine”, it is better to introduce a role “hasNationality” between Cuisine and Nationality.

The ontology has been designed in several steps. The initial design was based on common sense reasoning about the domain, incorporating ideas from the spatial ontology of UbisWorld[9] (the SpatialPurpose concepts) and the taxonomy of medical specialisations from the Bremen city portal[10]. Moreover, the choice of names for the classes of the ontology was inspired by the OSM tags; this helps in generating the ontology mapping automatically.

One main concept in the ontology is *Activity*, which implicitly means a location where an activity takes place. The selected activities are dependent on the particular scenario chosen for the application. We have currently concentrated on daily life activities and tourism. Notice however that the approach is flexible and the ontology can be easily adapted if an alternative scenario is chosen (e.g. business activities). Related activities are grouped and form subclasses of a certain activity type; this provides the advantage that the search can be done in such a way that all found locations can be displayed with a single search.

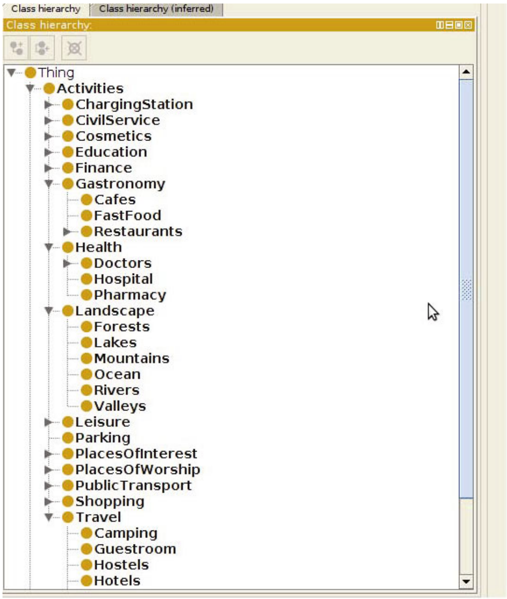
Locations can have associated addresses or opening hours. We decided to model these as abstract concepts *OpeningHours* and *Address*, and to make the analysis of the query at the string level. In the case of the opening hours, for example, they are usually stored in the OSM database as strings containing assignments of time intervals to the days of the week. The query given by the user will also be retrieved as a set of time intervals and the test whether a certain location satisfies a certain time restriction is done using Allen’s interval calculus [2], in particular, using the relations “contains” and “overlaps” between intervals.

To illustrate the ontology design, we present in the following the Restaurant concept.

The restaurants can be categorized according to their cuisine (see Fig. 7, where concepts are represented as discs, roles as arrows and subconcepts as dotted lines). The cuisine is either a food speciality, like pizza or seafood, or

[9] `http://ubisworld.ai.cs.uni-sb.de/index.php`

[10] `http://www.bremen.de/gesundheit_und_soziales/aerztesuche`



```
Thing
  Activities
    ChargingStation
    CivilService
    Cosmetics
    Education
    Finance
    Gastronomy
      Cafes
      FastFood
      Restaurants
    Health
      Doctors
      Hospital
      Pharmacy
    Landscape
      Forests
      Lakes
      Mountains
      Ocean
      Rivers
      Valleys
    Leisure
    Parking
    PlacesOfInterest
    PlacesOfWorship
    PublicTransport
    Shopping
    Travel
      Camping
      Guestroom
      Hostels
      Hotels
```

Fig. 6. Ontology of activities

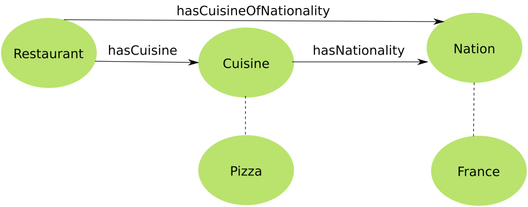


Fig. 7. Restaurants with cuisine

nation-specific, say French or Italian specialities. We model this by introducing concepts Restaurant, Cuisine and Nation, together with corresponding roles: a restaurant can have a cuisine and a cuisine can have a nationality. For cuisines and nationalities, we introduce the corresponding subconcepts. Strictly speaking, France should be an individual of the class Nation, but for keeping the symmetry with the ontology of tags, we prefer to introduce it as a singleton class. Notice that French restaurants could be expressed in OWL as the concept

```
Restaurant and hasCuisine some (hasNationality some France)
```

However, a conceptual design problem in the structure of tags in OpenStreetMap requires the presence of the role $hasCuisineOfNationality$ as composition of $hasCuisine$ and $hasNationality$ and thus French restaurants are equally represented as

```
Restaurant and hasCuisineOfNationality some France
```

The entire ontology has been subject to an evolving process. As mentioned, the ontology of spatially located activities will be related to an ontology of OpenStreetMap tags. Some tags may refer to activities that were not taken into account when designing the initial ontology. Of course, the challenge is to add such new concepts to the ontology of activities in an automatic manner. We will address this in the next section. Moreover, other data sources could be plugged in, like e.g. Google Maps. The situation repeats: the new database may contain data that was previously not considered to relate to interesting activities, but has since become relevant in the new context. It is then sensible to add a corresponding concept also at the abstract level, namely the ontology of activities.

6 An Ontology of OpenStreetMap Tags

OpenStreetMap's internal files are lists of nodes, ways and relations, which can be *tagged* with information about the map element. The convention is that any user is free to introduce his own tag, but it is recommended to use existing tags and only have new ones if they are not already covered by the existing ones. The tags of the map elements are represented as pairs (*key*, *value*) and an element of the map may have multiple tags (see Figure 8 for the example of a OSM node with its tags in an XML representation. This format has been developed by the OSM community. The listed tags vary from node to node).

```
<node id="834034642"
   lat="53.0871310" lon="8.8091071"
   version="7" changeset="6027662"
   user="Kerridge" uid="324245"
   timestamp="2010-10-13T09:51:39Z">
    <tag k="addr:city" v="Bremen" />
    <tag k="addr:country" v="DE" />
    <tag k="addr:housenumber" v="20" />
    <tag k="addr:postcode" v="28215" />
    <tag k="addr:street" v="Theodor-Heuss-Allee" />
    <tag k="amenity" v="charging_station" />
    <tag k="name" v="Elektrotankstelle swb" />
    <tag k="note" v="telephone reservation necessary" />
    <tag k="opening_hours" v="Mo-Fr  6:00-18:00; Sa off; Su off" />
    <tag k="operator" v="swb" />
    <tag k="phone" v="+49 421 3593186" />
  </node>
```

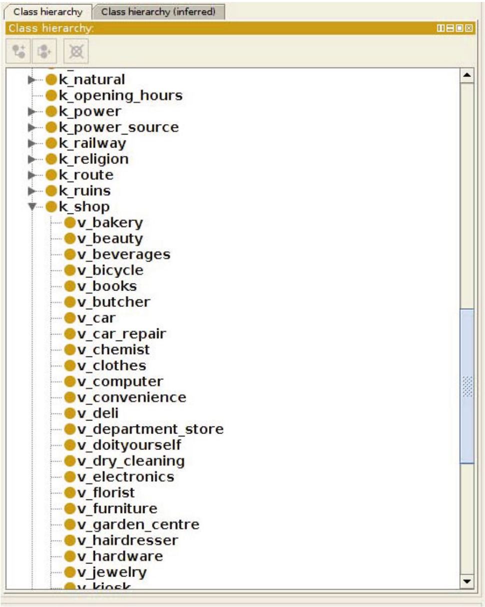
Fig. 8. A node in an OSM file

The purpose of the ontology of tags was to stay as close as possible to the structure of the OSM files in order to facilitate database querying. This means that we do not try to correct any possible conceptual mistakes in the taxonomy of OSM tags, but rather have it reflected faithfully in the structure of the ontology.

When designing the ontology, it makes sense to decompose the tags into a hierarchy according to the keys: the key becomes a superconcept of its values. We have followed this approach whenever the value was an OSM constant rather then a string/numeral. Since it is possible that a key and a value have the same name whilst the names of the concepts are required to be unique in OWL (OSM has “station” as value of the key “railway” but also a key named “station”), we decided to prefix all keys with “k_” and all values with “v_”, e.g.:

```
k = "amenity" v = "charging_station"
```

would introduce a concept “k_amenity” with a subconcept “v_charging_station”. Moreover, another problem is that some values are subclasses of more than one key. E.g. “v_no” is a subclass of “k_smoking” but also of “k_smoking_outside”.[11] In this case, we extended the value to “v_smoking_k_no”. Another design decision was to take into account tag dependencies. For example, when a node is tagged with `k = "amenity" v = "restaurant"` it is possible (but not mandatory) that the cuisine is also tagged: `k = "cuisine" v = "seafood"`. In such cases, we introduce a role *hasCuisine* with “v_restaurant” in domain (it is also possible that “v_fast_food” is tagged with “k_cuisine”) and range “k_cuisine” in order to be able to select only those restaurants with a certain cuisine.



- k_natural
- k_opening_hours
- k_power
- k_power_source
- k_railway
- k_religion
- k_route
- k_ruins
- k_shop
  - v_bakery
  - v_beauty
  - v_beverages
  - v_bicycle
  - v_books
  - v_butcher
  - v_car
  - v_car_repair
  - v_chemist
  - v_clothes
  - v_computer
  - v_convenience
  - v_deli
  - v_department_store
  - v_doityourself
  - v_dry_cleaning
  - v_electronics
  - v_florist
  - v_furniture
  - v_garden_centre
  - v_hairdresser
  - v_hardware
  - v_jewelry
  - v_kiosk

Fig. 9. Ontology of OSM tags

[11] We maintain our design uniform, so “v_no” must be a concept; other choices would also be available.

Recalling the example of French restaurants of the previous section, notice that nation specific cuisines are added directly as subconcepts of "k_cuisine"—see Fig. 10. This is conceptually a mistake in the design of OpenStreetMap's tags, which we here reflect in the ontology which is meant to be very close to the OSM tags.

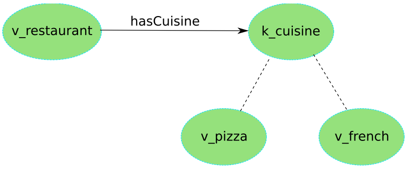


Fig. 10. Restaurants with cuisine

In order to create a realistic ontology of OSM tags, one faces the problem of an open project where everyone is allowed to contribute—which is also an OSM strength. This has the effect that the data source can be regarded as dynamic, not only at the level of entries, but also at the level of tags. In the OSM wiki page[12], there exists a list of tags, but this list does not reflect the status quo of the actual OSM databases. Some tags which are in the wiki page are not yet tagged by the community, some tags which were abolished through discussion in the wiki or the mailing lists are still used by the mappers. Therefore, the wiki provides only an overview of the available tags; to have a more realistic estimation, one should use websites like Taginfo[13], where the OSM data of the whole world is searched and a list of tags in use, sorted by the number of occurrences, is provided as a result. Of course, this list will also contain spelling errors or falsely used tags. The most straightforward solution here is to consider relevant those tags that have a certain, high occurrence in the database, using the list provided by Taginfo. This strategy could result in a limitation using a certain percentage (e.g., all values with, say, more than 0.3 % occurrence rate for the respective key are included), but this approach fails to capture all interesting values in the cases where some keys appear with a far higher occurrence and thus the percentage of important values is low. Also, some keys have far more values (e.g., amenity with 7714 values in use according to Taginfo) than others (e.g., smoking with 22 values), so that the percentage of each value naturally is quite low, which is another point against a certain percentage as a limit for inclusion. This is why a limitation based on the absolute occurrence of a value makes more sense. In our case, we decided to select all values which occur more than 100 times in the database. Spelling errors are thus excluded as well (there is never 100 times the same mistake), and still all relevant values will be in the database. Theoretically, this threshold could be exceeded by mistakes created during automatic tagging procedures. In reality, there is no evidence in Taginfo that this is the case. It is either prevented by the professionalism of those using automatic tagging, or mistakes of such quantity are quickly noticed by the community and repaired.

[12] http://wiki.openstreetmap.org/wiki/Map_features

[13] http://taginfo.openstreetmap.de/

After this procedure, we added the tags that are in the wiki but not covered through our search of Taginfo. This guarantees that we also include tags which are not yet used by the mappers, but in the future shall be implemented or will replace other tags. To keep this ontology of tags up-to-date, one option would be to make it available to the OSM community. People creating new tags could include them themselves into the ontology as well. Another option is automation, e.g., programs searching regularly through Taginfo for new tags.

7 Mapping the Ontologies

The connection between the two ontologies is bi-directional. In one direction, we map atomic concepts and roles from the ontology of activities to concepts and roles in the ontology of tags. This is the first step towards assigning the elements of the ontology of activities to queries over the database, and will be completed in the next section.

We allow the possibility that the ontology of activities contains concepts which are not related with OpenStreetMap tags. The reason for this is that it is possible to extend and complement the tool with a similar construction for other geographical database systems—indeed, the integration with Google Maps is currently in progress, and this means that some activities with no counterpart in OpenStreetMap may still be found using another database. Thus, the mapping we obtain is partial, and we can see this as having a sub-ontology of activities which is then mapped totally to the ontology of OSM tags.

Since the number of concepts and roles is quite large, providing such a mapping manually would be a very tedious process. We can, however, use an ontology matching tool to obtain a list of pairs of concepts that are in correspondence. This approach is very effective—with the ontology matcher Falcon [8], the degree of automation reaches 80%. This means that the user is still required to verify and confirm the matches produced with the tool, and possibly introduce new matchings between concepts that were not identified by the tool's analysis.

In the other direction concerning the connection of the ontologies, the evolving process for the ontology of activities has to be considered. The social character of OSM makes the tags subject to continuous change: new tags are added frequently and they are often modifying. An example particularly relevant for the topic of the paper is the tag for charging stations for electric cars: initially, a charging station was tagged as

```
k="amenity" v="fuel"
```

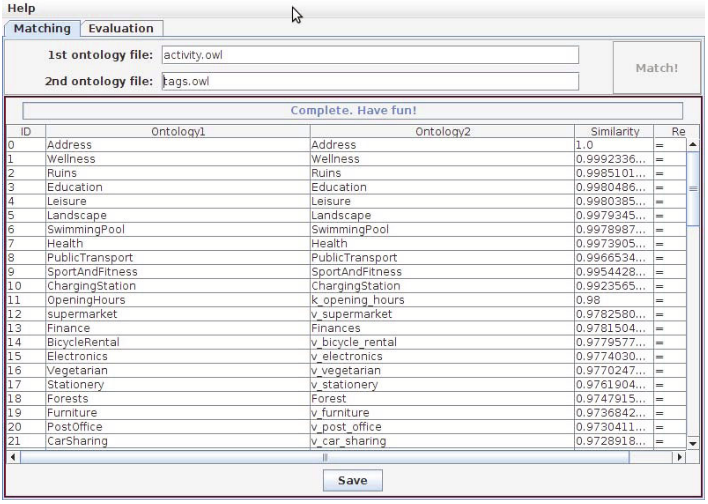
like any fuel station and with a supplementary tag

```
k="fuel:electricity" v="yes"
```

This has later evolved into introducing a distinguished value for amenity:

```
k="amenity" v="charging_station"
```

but the two ways of tagging charging station still coexist. The dynamic character of the OSM database should be reflected in our tool as well. As more locations



1st ontology file: activity.owl

2nd ontology file: tags.owl

Complete. Have fun!

ID	Ontology1	Ontology2	Similarity	Re
0	Address	Address	1.0	=
1	Wellness	Wellness	0.9992336...	=
2	Ruins	Ruins	0.9985101...	=
3	Education	Education	0.9980486...	=
4	Leisure	Leisure	0.9980385...	=
5	Landscape	Landscape	0.9979345...	=
6	SwimmingPool	SwimmingPool	0.9978987...	=
7	Health	Health	0.9973905...	=
8	PublicTransport	PublicTransport	0.9966534...	=
9	SportAndFitness	SportAndFitness	0.9954428...	=
10	ChargingStation	ChargingStation	0.9923565...	=
11	OpeningHours	k_opening_hours	0.98	=
12	supermarket	v_supermarket	0.9782580...	=
13	Finance	Finances	0.9781504...	=
14	BicycleRental	v_bicycle_rental	0.9779577...	=
15	Electronics	v_electronics	0.9774030...	=
16	Vegetarian	v_vegetarian	0.9770247...	=
17	Stationery	v_stationery	0.9761904...	=
18	Forests	Forest	0.9747915...	=
19	Furniture	v_furniture	0.9736842...	=
20	PostOffice	v_post_office	0.9730411...	=
21	CarSharing	v_car_sharing	0.9728918...	=

Fig. 11. Matching ontologies with Falcon

on a map are being tagged, it is reasonable to expect that tags that did not pass the criteria for being selected in the ontology of tags now become relevant and should be therefore included. We make use of this process to make the ontology of activities evolve as well: if a certain tag denotes an activity (this could be verified semi-automatically by inspecting its superconcept), it can be added to the ontology of activities.

7.1 The Heterogeneous Tool Set

The Heterogeneous Tool Set Hets [13] is a heterogeneous specification and proof management tool, providing support for a multitude of logics (including OWL in Manchester syntax) and interfacing various logic specific tools like theorem provers, consistency checkers etc. It relies on a heterogeneous specification language with an origin in CASL [14], a specification language developed within the IFIP working group 1.3 "Foundations of System Specifications", and is a de facto standard in the area of software specification. This is particularly relevant for ontology specification as this language can be employed for providing support for modularisation and structuring (see [11] for a detailed analysis). Also, when discharging a proof obligation in the system, if proof support is not directly available, a prover can be used by "borrowing" from another logic along a suitable translation of logics.

In the context of our application, Hets can be employed to verify semantic correctness of the ontology mapping produced by a matching. This means that the sentences of the sub-ontology of activities identified by the matching tool should translate along the mapping to logical consequences of the ontology of tags. The corresponding Hets specification can be obtained in an automatic manner: consider that ACTIVITIES is a named specification containing the initial ontology of activities and TAGS contains the ontology of tags.[14] The matching procedure returns a list of pairs having component concepts in the ontology of activities which have a corresponding match in the ontology of tags. This fits in exactly with the HETS syntax: ontology mappings are written as *symbol maps*, i.e., for a symbol of the source ontology, one must give its corresponding symbol along the mapping. Also, the source and target ontologies must be explicitly given; while the target specification is simply TAGS, the source specification is obtained from ACTIVITIES by *revealing* the symbols in the sub-ontology, an operation which hides the remaining ones. This finally gives a complete HETS specification of the ontology mapping and the tool can be used for verifying correctness.

The correctness of the mapping can be verified using HETS and the provers Pellet[15], Fact++[16] or also first-order provers like SPASS [20]. A Protégé [17] plugin for manipulating HETS-OWL specifications is under development.[18] In the current state of the ontologies, discharging the proof obligations is relatively simple. The added value of using a formal verification method for the view becomes visible in the presence of subsumptions implying more complex terms. Since such terms could be introduced by changes in the database or usage of another data source, we preferred to include this step as part of the tool methodology.

8 Ontology-Based Data Access

Ontology-based data access is a data integration methodology which separates the 'knowledge' about data from reasoning about it. This is achieved by providing an abstract representation of the application domain with the help of an ontology, a schema of the sources where the real data is stored, together with a mapping between the elements of the ontology and those of the data schema. Typically, the schema of the data is assumed to be a relational database schema, and the mapping provides a query in the database for each concept and each role of the ontology. The advantage of this approach is that we can use the knowledge base constituted by the TBox and the ABox sentences of the ontology to derive information about the data which is not present in the database, using query rewriting.

The data integration management component of our system follows the principles of OBDA: the domain of interest—spatially located activities—is modelled as an ontology, the OpenStreetMap data is stored in a database, and the concepts of the ontology are related to queries in the database.

For the representation of and access to ontologies within the Ruby on Rails framework, we have developed a new library, Rails-OWL. Since OWL is represented in XML, our library is based on the existing library REXML[19] for

[14] Notice that here ontologies are regarded as logical theories in the OWL logic.

[15] http://clarkparsia.com/pellet/

[16] http://owl.man.ac.uk/factplusplus/

[17] http://protege.stanford.edu/

[18] https://github.com/pyneo/protege-hets

[19] http://www.germane-software.com/software/rexml/

reading in XML documents. Rails-OWL represents an OWL ontology in the Rails database. This allows programmers to easily and flexibly access ontologies in a way similar to the access of the geodata. Fig. 12 gives an overview of the different classes (which by the ActiveRecord framework of Rails simultaneously are database tables) used for representing ontologies, their classes and mappings between these. A "simple subclass relation" is one between named concepts, while in general, it can be postulated between arbitrary OWL class terms.

database table	represented contents
Ontology	ontologies
OntologyClass	classes (of various ontologies)
OntologySubclass	simple subclass relations
OntologyClassProperty	subclass (and other) relations
OntologyRole	roles (of various ontologies)
OntologyRoleProperty	role relations
OntologyMapping	ontology mappings
OntologyMappingElement	mapped pairs (of various mappings)

Fig. 12. Classes for representing OWL ontologies in Ruby on Rails

In ontology-based data access, usually, one SQL query per ontology class is designed manually, and this is used for the database interpretation of ontology terms, implemented by query rewriting. In case of OpenStreetMap, we would need to design dozens of such SQL queries, which is a tedious process. Instead, we use the OSM tag ontology, which is tailored towards the OSM database in such a way that the relation between classes in the OSM tag ontology and the OSM database is generic: since the basic classes directly correspond to keys and values of OSM tags, the corresponding SQL queries are simple, and this is then used for query rewriting of more complex class terms. This query rewriting is implemented in Rails-OWL easily, because classes, roles and such are first-class citizens. The involved OSM tables are shown in Fig. 13.

database table	represented contents
Node	geographical location with coordinates
NodeTag	tags for nodes with key and values
Way	polyline
WayNode	incidence relation between nodes and ways
WayTag	tags for way with key and values

Fig. 13. Classes used by OpenStreetMap for representing Geodata in Ruby on Rails

9 Conclusion and Future Work

We have presented an ontology-based tool prototype for searching locations for specific activities in OpenStreetMap. We have here concentrated on presenting the architecture and the general ideas underlying our tool, and could only sketch some of the technical details. The focus is not primarily on locations as such, but rather on (located) activities. Here, the central ontology of spatially located activities is subject to evolution due to the continuous development of the OpenStreetMap databases, and we have introduced an intermediate ontology of data source representation terms to facilitate querying.

Future work will enhance the ontology-based querying with a more sophisticated ontology navigation and query refinement, see [6] for an overview of existing approaches. Also, the searching of activities will be complemented with an activity-oriented route planning (such that Scenarios 3 and 4 in Section 2 will be supported by our tool as well). Our main topic of interest is electric mobility which requires special route-finding algorithms that take into account the energy consumption and the (projected) battery status at a given destination point. We also intend to complement this with a radius visualisation of the area within battery reach.

Acknowledgments

This work has been supported by the German Research Foundation (DFG) and Project I1-[OntoSpace] of the SFB/TR 8 “Spatial Cognition”. We thank C. Maria Keet for proposing the name DO-ROAM and John Bateman and the anonymous referees for useful feedback. Moreover, Erwin R. Catesbeiana pointed out user queries that lead to no results. We also thank Shailen Sobhee and Shishir Gautam for implementation work.

References

1. Adabala, N., Toyama, K.: Purpose-driven navigation. In: Rodríguez, M.A., et al. (eds.) [18], pp. 227–233
2. Allen, J.F.: Maintaining knowledge about temporal intervals. Communications of the ACM 26, 832–843 (1983)
3. Bateman, J.A., Hois, J., Ross, R.J., Tenbrink, T.: A linguistic ontology of space for natural language processing. Artificial Intelligence 174(14), 1027–1071 (2010)
4. Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Poggi, A., Rodriguez-Muro, M., Rosati, R., Ruzzi, M., Savo, D.F.: The MASTRO system for ontology-based data access. Semantic Web Journal (2011) (forthcoming)
5. Fellbaum, C. (ed.): WordNet: An electronic lexical database. The MIT Press, Cambridge (1998)
6. Hoang, H.H., Min Tjoa, A.: The state of the art of ontology-based query systems: A comparison of existing approaches. In: Proc. of ICOCI 2006 (2006)
7. Horridge, M., Drummond, N., Goodwin, J., Rector, A., Stevens, R., Wang, H.: The Manchester OWL Syntax. In: OWL: Experiences and Directions (2006)
8. Hu, W., Qu, Y.: Falcon-AO: A practical ontology matching system. In: Proc. of WWW 2007, pp. 237–239 (2008)
9. Hübner, S., Spittel, R., Visser, U., Vögele, T.J.: Ontology-based search for interactive digital maps. IEEE Intelligent Systems 19(3), 80–86 (2004)

10. Kuhn, W.: Ontologies in support of activities in geographical space. International Journal of Geographical Information Science 15(7), 613–631 (2001)
11. Kutz, O., Mossakowski, T., Lücke, D.: Carnap, Goguen, and the Hyperontologies: Logical Pluralism and Heterogeneous Structuring in Ontology Design. Logica Universalis 4(2), 255–333 (2010); Special Issue on 'Is Logic Universal?'
12. Mizen, H., Dolbear, C., Hart, G.: Ontology ontogeny: Understanding how an ontology is created and developed. In: Rodríguez, M.A., et al. (eds.) [18], pp. 15–29
13. Mossakowski, T., Maeder, C., Lüttich, K.: The Heterogeneous Tool Set. In: Grumberg, O., Huth, M. (eds.) TACAS 2007. LNCS, vol. 4424, pp. 519–522. Springer, Heidelberg (2007)
14. Mosses, P.D.: CASL Reference Manual. LNCS, vol. 2960. Springer, Heidelberg (2004)
15. Poggi, A., Lembo, D., Calvanese, D., De Giacomo, G., Lenzerini, M., Rosati, R.: Linking data to ontologies. J. on Data Semantics X, 133–173 (2008)
16. Poggi, A., Rodriguez-Muro, M., Ruzzi, M.: Ontology-based database access with DIG-Mastro and the OBDA Plugin for Protégé. In: Patel-Schneider (ed.) Proc. of the 4th Int. Workshop on OWL: Experiences and Directions (OWLED 2008 DC), vol. 496, CEUR-WS (2008)
17. Raubal, M., Kuhn, W.: Ontology-based task simulation. Spatial Cognition & Computation 4(1), 15–37 (2004)
18. Rodríguez, M.A., Cruz, I.F., Egenhofer, M.J., Levashkin, S. (eds.): GeoS 2005. LNCS, vol. 3799. Springer, Heidelberg (2005)
19. Seifert, I.: Region-based model of tour planning applied to interactive tour generation. In: Jacko, J.A. (ed.) HCI 2007. LNCS, vol. 4552, pp. 499–507. Springer, Heidelberg (2007)
20. Weidenbach, C., Brahm, U., Hillenbrand, T., Keen, E., Theobald, C., Topic, D.: SPASS version 2.0. In: Voronkov, A. (ed.) CADE 2002. LNCS (LNAI), vol. 2392, pp. 275–279. Springer, Heidelberg (2002)

Urban Area Characterization Based on Semantics of Crowd Activities in Twitter

Shoko Wakamiya[1], Ryong Lee[2], and Kazutoshi Sumiya[2]

[1] Graduate School of Human Science and Environment, University of Hyogo, Japan
[2] School of Human Science and Environment, University of Hyogo, Japan
nd09a025@stshse.u-hyogo.ac.jp,
{leeryong,sumiya}@shse.u-hyogo.ac.jp

Abstract. It is essential to characterize geographic regions in order to make various geographic decisions. These regions can be characterized from various perspectives such as the physical appearance of a town. In this paper, as a novel approach to characterize geographic regions, we focus on the daily lifestyle patterns of crowds via location-based social networking sites in urban areas. For this purpose, we propose a novel method to characterize urban areas using Twitter, the most representative microblogging site. In order to grasp images of a city by social network based crowds, we define the geographic regularity of the region using daily crowd activity patterns; for instance, the number of tweets, through the number of users, and the movement of the crowds. We also analyze the changing patterns of geographic regularity with time and categorize clustered urban types by tracking common patterns among the regions. Finally, we present examples of several urban types through the observation of experimentally extracted patterns of crowd behavior in actual urban areas.

Keywords: Urban Characteristics, Microblogs, Geographical Regularity.

1 Introduction

Characterizing urban areas is essential for making various geographic decisions. For example, a family planning to move to a city may want to look for the ideal location by considering the region from various aspects such as safety, amenities, landscape, age-group, living levels of the inhabitants, and traffic congestion. In fact, geographic characteristics can be classified from various perspectives: by geographic shape or diverse physical objects such as streets or landmarks, or by cultural and structural aspects such as residential, commercial, and industrial districts. These two different views have been well studied in many research fields. Kevin A. Lynch's seminal contribution in his book titled "The Image of the City" [10] focused on how people perceive their living space using five fundamental elements of a city, that is, paths, edges, districts, nodes, and landmarks. Based on these elements, Lynch thought that people could characterize their living space within the scenery of a city to picture themselves working and living there. In another remarkable work describing a way to characterize urban areas, Tezuka et al. [14] utilized Web contents to extract frequently

C. Claramunt, S. Levashkin, and M. Bertolotto (Eds.): GeoS 2011, LNCS 6631, pp. 108–123, 2011.

mentioned geographic objects and their roles on the Web by analyzing linguistic patterns related to the names of the landmarks. In this work, the Web contents were utilized as a mirror of the crowds' minds to the real world.

In the present work, we challenge to explore an uncharted realm of human geography by utilizing current social networks. Actually, due to recent advances in location-aware social networking sites, represented by Facebook[1] or Twitter [15], crowds share updates in near real time, often recording their opinions and life logs. Interestingly, on Twitter, numerous location-based messages (called tweets) are being written by many people beyond generations not only the young but also the old; hence, tweets can be considered as a geo-social database of opinions of a large proportion of the population, indicating behavior and lifestyles relevant to those living in a particular urban space.

In this paper, we propose a novel method to characterize urban areas by extracting some common patterns of crowd activity on Twitter. In order to extract such patterns, we focused on geo-tagged tweets; specifically, the number of tweets written in a specific urban geographic area, the number of users, and the movement of the crowds. Based on these three indicators, we constructed a daily geographic regularity for a region for four 6-h time periods to represent the region's usual status. Then, we clustered regions considering similar change patterns of 6-h regularities. Finally, for each cluster showing common crowd activity patterns, we analyzed the types of commonly found districts to determine and understand the most frequently occurring social phenomena in specific urban areas.

The remainder of this paper is organized as follows: Section 2 describes the outline of our research model and reviews some related work. Section 3 provides the details of the overall process, focusing on estimation of geographical regularities and clustering of urban areas in which similar crowd behavior patterns are observed. Section 4 illustrates the experiment conducted using a real dataset collected from Twitter. Section 5 concludes this paper with a brief description of future work.

2 Twitter-Based Urban Area Characterization

2.1 Our Motivation

The manners in which people use urban spaces are an important aspect in determining that space's urban characteristics such as the roles or functions which an urban area provides, or an urban area's attractiveness to the public. Of course, the roles or functions of urban areas can be described by surveying the types of districts from a lot of residents or by observing the appearance of city [10]. However, the appearance of a city is not enough to determine its urban characteristics. Instead, our approach focuses on crowd activities, which are much more informative and dynamic, but are currently an unexplored resource in describing crowd-centric urban spaces. Therefore, we expect that utilizing crowd activities for the characterization of urban areas is to be a valuable undertaking.

Furthermore, as a source from which to gather such crowd activity data, we chose the popular microblogging site, Twitter, where a large number of people around the

[1] Facebook: `http://www.facebook.com/`

world share updates and information about their whereabouts. Nowadays, according to Pew Internet's survey [13], the user groups are expanding to the older ages as well as the young. Indeed, Twitter has shown significance as an advanced location-based social network site; specifically, because today's smartphone technology has location measuring functionality, it is easy to write and share geo-tagged tweets. Therefore, we can consider the utilization of such geo-tagged tweets, which, in a sense, are a life log of each user, indicating when and where a user exists and what s/he is currently doing. Accordingly, we can construct a socio-geographic database using Twitter and analyze crowd lifestyle patterns, enabling us to practically extract urban characteristics. For example, Fig. 1 demonstrates the discovering of geographic characteristics based on "crowd behavior." In the figure, the map drawn in the center shows an urban area where many towns are characterized by the lifestyle patterns of the crowds in each region, such as the various places for working, drinking, eating, living, shopping, sightseeing, learning, etc. In the case of working place, we can observe that crowd lifestyles would routinely repeat similar patterns depending on their office hours of weekdays. Furthermore, in this paper, in order to characterize geographic regions by a quantitative approach based on enormous number of crowd life logs obtained from Twitter, we are more interested in the crowd activities reflected on Twitter rather than the actual tweet messages, while the tweet messages would be helpful in the later stage of understanding characterized urban areas.

In this section, we highlight our research model for realizing urban area characterization utilizing the collective experiences of local crowds sourced from Twitter. In particular, we present an overview of how we can exploit geo-tagged tweets to elicit crowd behaviors, which are used for estimating geographical regularity. Lastly, we review some related work that has previously attempted to conduct socio-geographic analytics using Twitter.

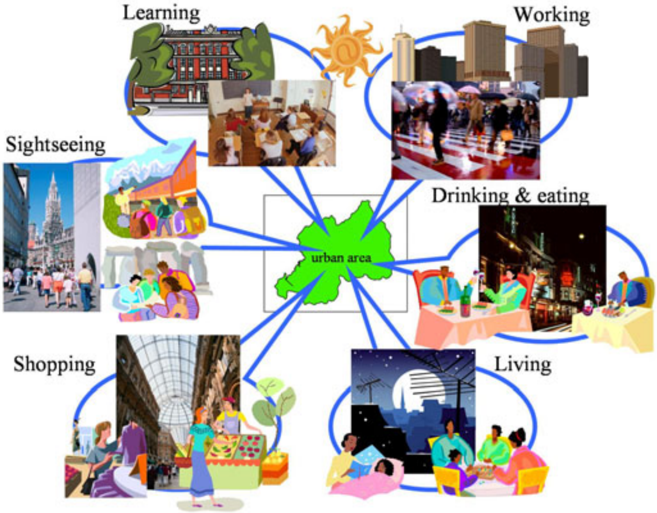
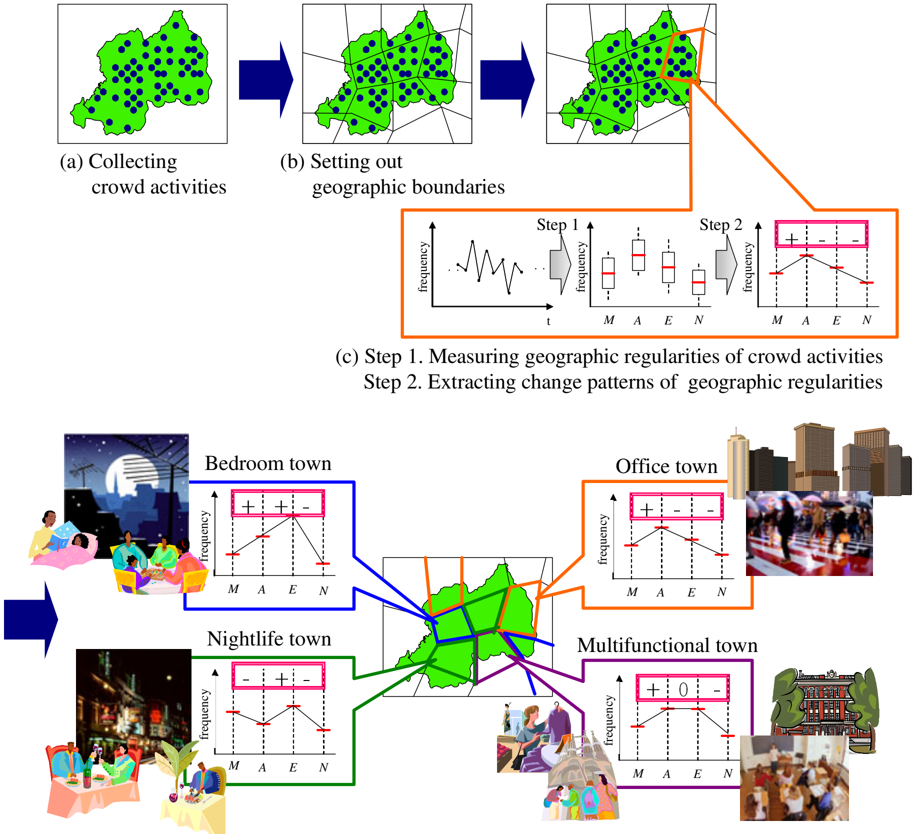


Fig. 1. Geographic characteristics based on crowd activities

2.2 Our Research Model

First, we present our research model for extracting urban characteristics using Twitter. As previously mentioned, we aim to reveal an urban area's features relevant to crowd activities. It is possible to obtain such data through the Open API in Twitter [16]. The tweets which we were able to collect from the site consisted of the four attributes of user id, written time, geographic location, and textual message. From this primitive form of data, we were able to identify the number of tweets occurring in a particular region, the number of people, and lastly, by referring to old logs, people's movement patterns. In the present work, we began with the process by gathering such geo-tagged tweets from the site, as shown in Fig. 2 (a).

Next, to investigate partial regions of the target area, we needed to properly partition our whole area of interest. Here, we established the partitions based on the distribution of tweets, as shown in Fig. 2 (b). We thought that this kind of partitioning



(a) Collecting crowd activities

(b) Setting out geographic boundaries

(c) Step 1. Measuring geographic regularities of crowd activities
Step 2. Extracting change patterns of geographic regularities

(d) Characterizing urban areas based on patterns of crowd activities

Fig. 2. Process of urban area characterization

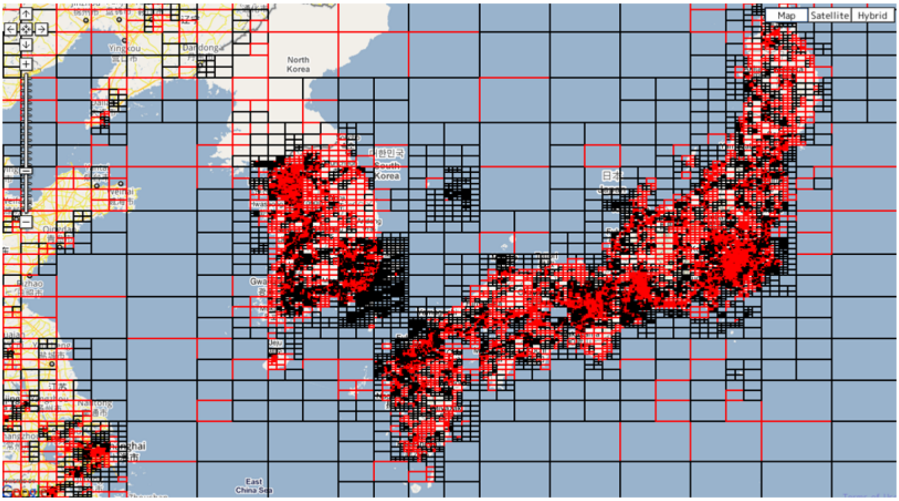


Fig. 3. Geographic distribution of tweets found in Japan and Korea

would well represent socio-graphic boundaries, strongly supported by crowd activities. Subsequently, we investigated crowd activity patterns in each partitioned region by estimating geographic regularities based on crowd activities extracted from the Twitter data we collected, as shown in Fig. 2 (c). Finally, we clustered urban areas and found significant types based on daily change patterns of geographic regularities and characterized urban areas and the meaning of each type of area based on the patterns: bedroom town, office town, nightlife town, and multifunctional town, as shown in Fig. 2 (d).

(1) Collecting Crowd Activities via Twitter

While it is possible to collect tweets from Twitter, it takes a considerable amount of effort to gather a significant number of geo-tagged tweets because of certain practical limitations: first, Twitter's open API solely supports the simplest near-by search by means of the specification of a center location and a radius. Furthermore, each query can only obtain a maximum of 1,500 tweets per week. To overcome these restrictions and perform periodic monitoring of any user-specified regions, we developed a geographical tweet gathering system [2] that could collect significant amounts of geo-tagged microblog data for a region of any size. For instance, Fig. 3 shows a map superimposed with geo-tagged tweets around Japan; it actually illustrates a quad-tree where the small colored cells represent densely populated regions.

(2) Establishing Socio-geographic Boundaries

Next, to characterize geographic regions in a given large area, we needed to determine how to partition the target region into sub-areas in order to examine the usual patterns of the crowds in those areas. In order to configure socio-geographic boundaries properly, we adopted a clustering-based space partition method that could reflect the geographical distribution of a dataset and deal with heterogeneous regions differently. Specifically, we adopted the K-means clustering method [5] based on the geographical occurrences of a dataset. Then, we regarded the K-partitioned regions over a map as socio-geographic boundaries, which were later formed by a Voronoi diagram using the centers of the clusters. Consequently, we were able to effectively perform the appropriate setting of socio-geographic boundaries for the target region by referring to the actual tweets' occurrences and focusing on the major hotspots.

(3) Estimating Geographical Regularity of a Crowd's Activity Pattern

To conduct the urban area characterization, it was necessary to determine the ordinary or usual patterns of local crowd activity. For this, we estimated the geographical regularity for socio-geographic boundaries during a certain time period. We examined these factors every 6-h by splitting a day into four equal time periods of morning, afternoon, evening, and night; we decided this temporal granularity on the basis of social crowd behaviors related to meal time. We defined three types of crowd activities for a region— (i) the total number of tweets happening, (ii) the number of distinct users, and (iii) the number of distinct moving users—and summarized them as boxplots [11].

(4) Characterizing Urban Areas Based on Geographical Regularities

The characteristics we present in this paper are sourced from the activity patterns of local users on Twitter. To describe urban characteristics using them, we created a simplified representation to depict temporal changes by computing the differences between two consecutive periods of time as $uc_T(g_i) = (m_1\text{-}m_0, m_2\text{-}m_1, m_3\text{-}m_2)$, where m_j means an estimated value at a period of time t_p to denote the urban characteristic of a geographic area g_i in terms of tweets. Again, we kept it simple by only looking up the method of change by transforming the representation to a symbol list such as $uc_T(g_i)$ = (+, ‐ , 0), which respectively increases from m_0 to m_1, decreases from m_1 to m_2, and indicates no change from m_2 to m_3. Indeed, the changes dynamically signified crowd activities. This process will later be described in detail.

2.3 Related Work

In order to characterize urban areas, a variety of research studies have been conducted utilizing social crowd-based sources. Vieira et al. [17] have proposed a Dense Area Discovery (DAD-MST) algorithm for automatically detecting dense areas using the ubiquitous infrastructure provided by a cell phone network. Kurashima et al. [7] have developed a Blog Map of Experiences, which can share personal experiences of tourists at specific locations and times extracted by using association rules from blog entries and present them visually; their method could characterize sightseeing spots by means of visitors' activities and evaluations. In addition, Moriya et al. [12] developed a system that estimates images, impressions, or the atmosphere perceived by bloggers of a region, and displayed the results on a digital map.

As studies conducted on the basis of geo-social tweets database, Fujisaka et al. [1] proposed a method for detecting geo-social events including unexpected events based on crowd moving pattern; aggregation and dispersion. Furthermore, Lee et al. [8, 9] presented to detect geo-social events by measuring crowd activities' regularity. These previous methods applied to a similar approach with our work, however, the difference is these work's goal aiming to detect geo-social events borrowing crowd power obtained from Twitter. On the other hand, our proposed work's purpose is to characterize urban areas on the basis of geo-social tweets database.

As a trend of recent social networking services, microblogs have received considerable publicity in not only as services for ordinary people but also as many academic and practical challenges. Specifically, Krishnamurthy et al. [6], Java et al. [4], Zhao et al. [19], and examined focusing on utilization of Twitter by making a relation to its impact on lifestyles and topical tendencies. Also, Iwaki et al. [3] have proposed a method for the discovery of useful topics from microblogs. These studies mostly dealt with content analyses of textual messages and the link structure of the followers of certain users.

3 Measuring Geographical Regularity of Urban Areas

In this section, we describe the details of regarding 1) how the socio-geographic boundaries were prepared to enable us to partition an observing geographic area reasonably, 2) how the geographic regularities for each urban area were established based on crowd behavior, and 3) how the urban areas were characterized.

3.1 Configuration of Socio-Geographic Boundaries

We initially needed to determine a set of urban areas from the geo-tagged dataset, which we had accumulated in advance using our geographical tweet monitoring system. Each tweet consists of attributes such as a user ID, time stamp, location (geo-tag) by latitude and longitude, and textual message. Here, the location can be extracted either from raw text form or in very precise location coordinates. Hence, in the former textual style, we needed to perform geo-coding to identify the exact coordinates by translating place names into the corresponding exact locations. We were able to solve the problem easily by using another mash-up service with Google Map's geo-coding service[2]. This useful conversion service enabled us to transform the place names to the precise coordinates directly. Thus, we could accurately determine when and where each tweet was written. Next, we set out geographic boundaries from the dataset which we clarified the geospatial and temporal occurrences of the tweets. In order to establish the boundaries, we simply classified geo-tagged tweets on the basis of latitudes and longitudes by K-means [5]. Then, we partitioned our target region of Japan into the same number of sub-regions, as illustrated in Fig. 4. We define these partitioned sub-regions as experimental areas which we attempt to figure out geographic regularity.

[2] The Google Geocoding API: `http://code.google.com/intl/en/apis/maps/documentation/geocoding/`

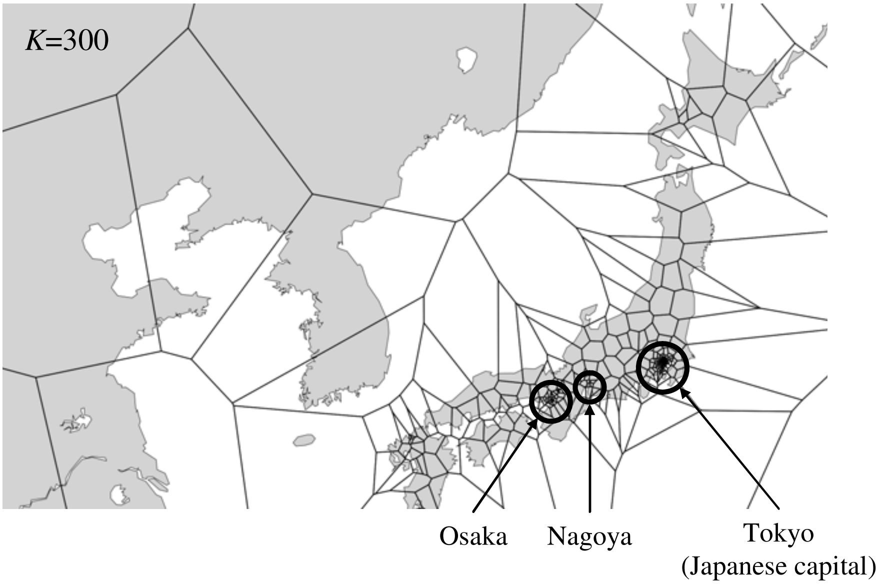


Fig. 4. Voronoi representation of socio-geographic boundaries based on tweets around Japan (constructed from K-means)

3.2 Measuring Geographical Regularity

Under the assumption that it is possible to grasp characteristics of urban areas from crowd activities, we present a method to summarize the usual patterns of these activities in a succinct representation. As previously mentioned, we denoted such patterns for socio-geographic boundaries as geographical regularity. Furthermore, we established each socio-geographic boundary's geographical regularity (*gr*) based on the following indicators:

1) *#Tweets*: The total number of tweets occurring inside of the socio-geographic boundaries within a specific period of time.

2) *#Crowd*: The total number of Twitter users found within the socio-geographic boundaries within a specific time period. In general, the in-equality *#Crowd* ≤ *#Tweets* is valid since any individual can write one or more tweets during the specific period of time.

3) *#MovCrowd*: The number of moving users related to socio-geographic boundaries within a specified period of time. In terms of partitioned socio-geographic boundaries, there are three types of moving user groups: a) Inner: a crowd in socio-geographic boundaries moves only inside the region without going outside of it; b) Incoming: there are some people coming from outside of the region; and c) Outgoing: conversely, some people move outside the boundaries. To simplify the cases as much as possible, we only considered the inner user moving groups.

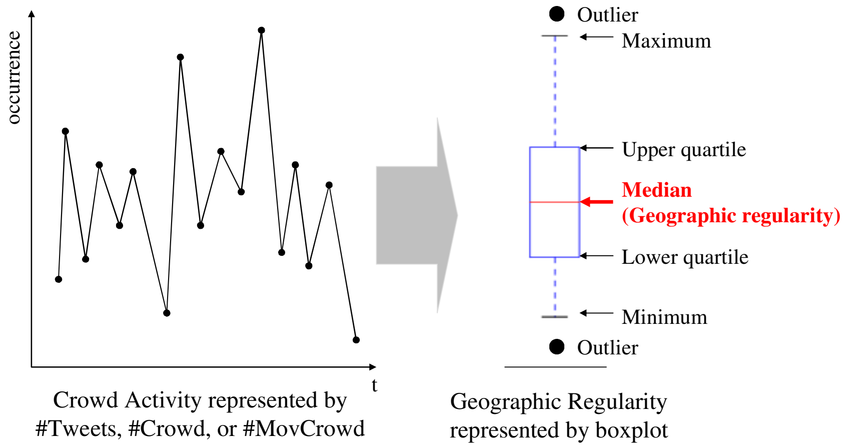

occurrence
t
Crowd Activity represented by #Tweets, #Crowd, or #MovCrowd
Outlier
Maximum
Upper quartile
Median (Geographic regularity)
Lower quartile
Minimum
Outlier
Geographic Regularity represented by boxplot

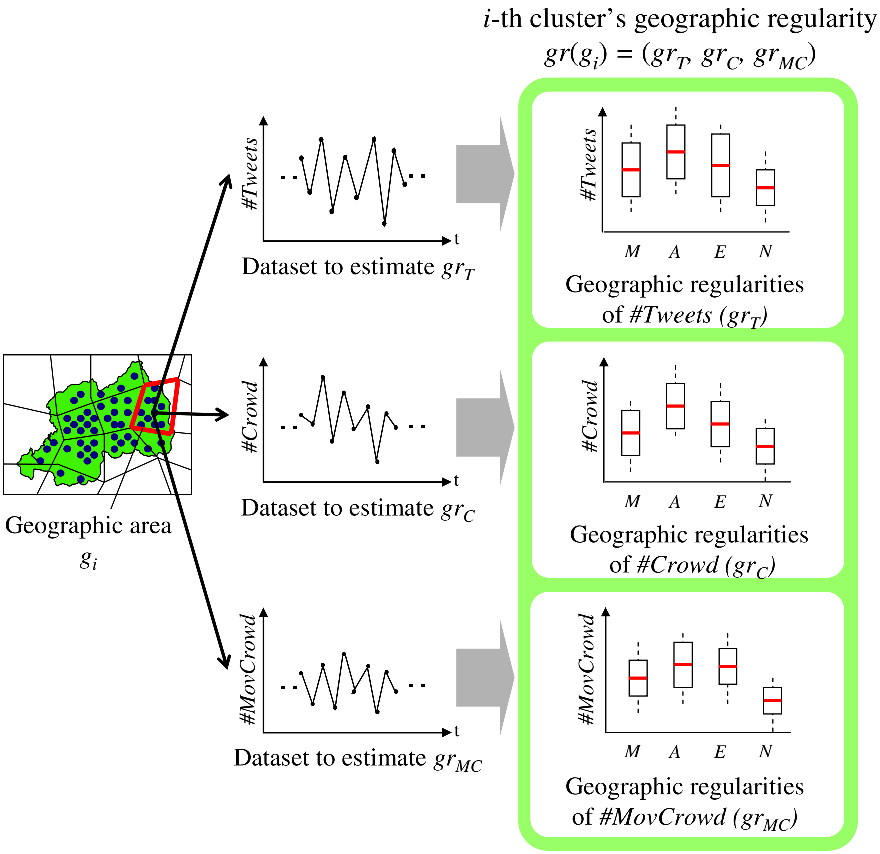
Fig. 5. Boxplot-based geographical regularity construction

Next, we derived the usual patterns of these three indicators from a long-term training dataset. For the sake of simplicity, we dealt with all the indicators in a statistical manner by using a boxplot [11], which is primarily used for explicitly visualizing a data distribution, as shown in Fig. 5. As depicted in Fig. 6, we built geographical regularities from the three indicators. That is, for spaces made by socio-geographic boundaries during a specific time period, gr_T, gr_C, and gr_{MC} represent the geographical regularities for *#Tweets*, *#Crowd*, and *#MovCrowd*, respectively. We considered that individual spaces could have a different quantity of patterns and, hence, established a geographic regularity for each socio-geographic area from these three indicators by the notation $gr(g_i) = (gr_T, gr_C, gr_{MC})$, where $g_i \in$ socio-geographic areas G.

3.3 Characterizing Urban Areas

We now describe a method to extract urban characteristics utilizing geographic regularity patterns. To this point, we have constructed the regularity of each geographic region using three succinct boxplots, as shown on the right side of Fig. 6. Furthermore, we divided a 24-h day into four 6-h periods: morning (M, 06:00–12:00), afternoon (A, 12:00–18:00), evening (E, 18:00–24:00), and night (N, 24:00–06:00). Thus, for each region, there would be three regularity expressions for the number of tweets, the crowds, and the number of moving crowds: $gr_T = (d_M, d_A, d_E, d_N)$, $gr_C = (d_M, d_A, d_E, d_N)$, and $gr_{MC} = (d_M, d_A, d_E, d_N)$, respectively. Here, d_p means a box plot range given by (min, max) and a median value for a period of time, p.

However, to find out common patterns among the usual types of regularity, which consist of four ranges, we must consider a much simpler expansion, since the ranges of each region can be different and their comparison can be complicated. In fact, the



i-th cluster's geographic regularity
$gr(g_i) = (gr_T, gr_C, gr_{MC})$
#Tweets
t
Dataset to estimate gr_T
#Crowd
Dataset to estimate gr_C
#MovCrowd
Dataset to estimate gr_{MC}
Geographic area
g_i
M A E N
Geographic regularities of #Tweets (gr_T)
Geographic regularities of #Crowd (gr_C)
Geographic regularities of #MovCrowd (gr_{MC})

Fig. 6. Estimation of geographical regularity

characteristics of our urban area of interest are about relative change patterns, that is, the number of crowds existing in a region can be different, but two or more regions can be similar in an increasing tendency, for example, from morning to afternoon.

Next, we were interested in finding the common patterns within the change patterns. Specifically, each pattern can be a symbol list of "+," "0," " — ," respectively, for increasing, staying, and decreasing, so that there can be many combinations such as (+, +, +) and (+, —, 0), as illustrated in Fig. 7. Lastly, we established an urban characteristic pattern by concatenating the three patterns. Thus, the possible combinations could be large (= 3^9) and the computational cost to find out the common partial or full-size patterns would be unbearable. In order to reduce such cost, we applied a Frequent Itemset Mining algorithm [18], statistically constructing it to the full size of the pattern (here, the length is set to 9).

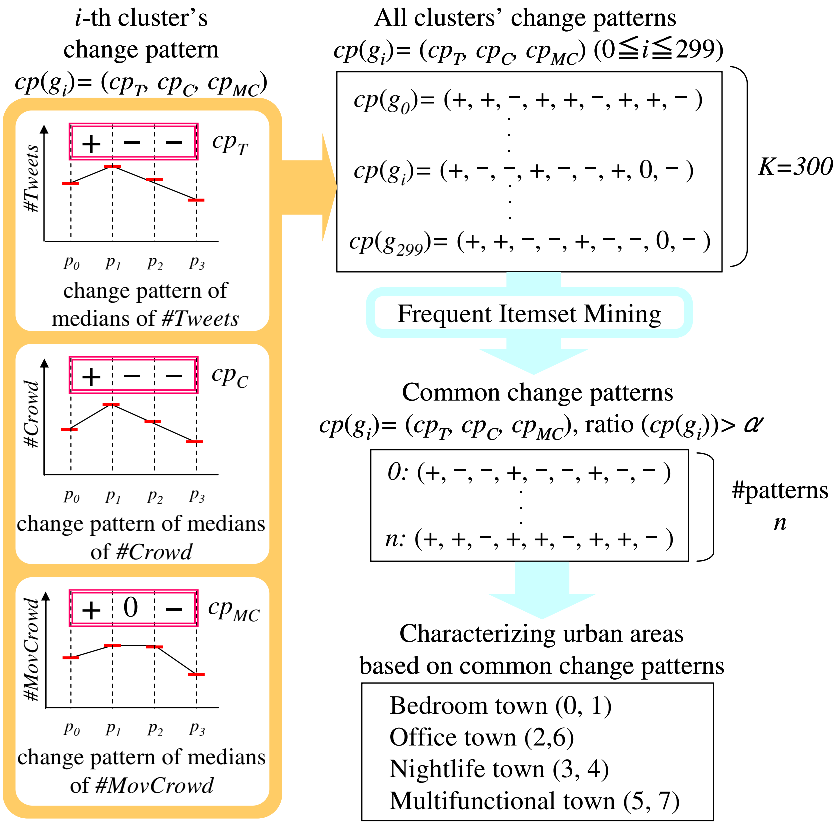

i-th cluster's change pattern
cp(g_i)= (cp_T, cp_C, cp_MC)
#Tweets
+ − −
cp_T
p0 p1 p2 p3
change pattern of medians of #Tweets
#Crowd
+ − −
cp_C
p0 p1 p2 p3
change pattern of medians of #Crowd
#MovCrowd
+ 0 −
cp_MC
p0 p1 p2 p3
change pattern of medians of #MovCrowd
All clusters' change patterns
cp(g_i)= (cp_T, cp_C, cp_MC) (0≦i≦299)
cp(g_0)= (+, +, −, +, +, −, +, +, −)
cp(g_i)= (+, −, −, +, −, −, +, 0, −)
cp(g_299)= (+, +, −, −, +, −, −, 0, −)
K=300
Frequent Itemset Mining
Common change patterns
cp(g_i)= (cp_T, cp_C, cp_MC), ratio (cp(g_i))> α
0: (+, −, −, +, −, −, +, −, −)
n: (+, +, −, +, +, −, +, +, −)
#patterns n
Characterizing urban areas based on common change patterns
Bedroom town (0, 1)
Office town (2,6)
Nightlife town (3, 4)
Multifunctional town (5, 7)

Fig. 7. Urban area characterization based on change patterns of geographic regularity

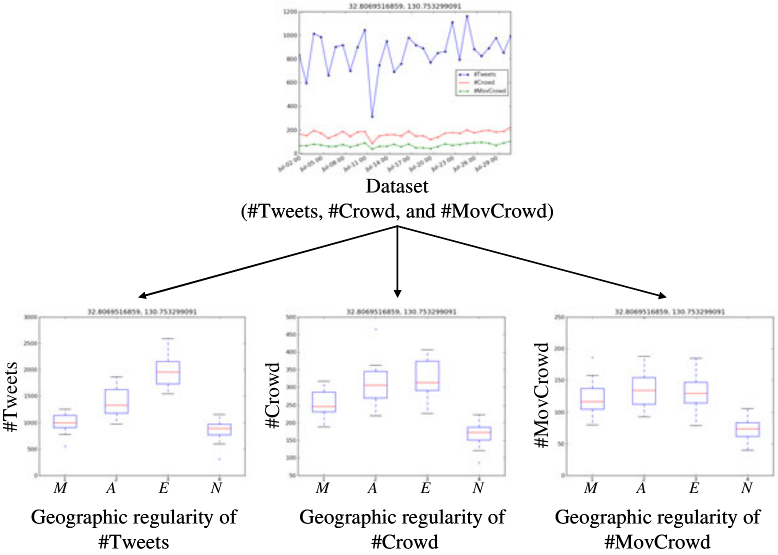
4 Experiments and Evaluation

4.1 Data Collection and Setting Out Geographic Boundaries

First, we obtained geo-tagged tweets for one month (2010/07/01–2010/07/30) around Japan with the latitude range [30.004609:45.767523] and the longitude range [116.27921:148.381348] using our geographical tweets collecting system as shown in Fig. 3. We gathered 11,632,750 geo-tagged tweets from 211,361 distinct users. However, in this experiment, we utilized only the tweets found in the areas of Japan chosen for the characterization of urban areas. Next, we constructed a set of socio-geographic boundaries using the obtained data. As explained in Section 3.1, we divided the target space by a K-means clustering method, using the data with a condition of $K = 300$. The results are presented in Fig. 4.

4.2 Estimating Geographic Regularity

We estimated the geographical regularities for clusters in terms of *#Tweets* (gr_T), *#Crowd* (gr_C), and *#MovCrowd* (gr_{MC}) for every fixed time slot; empirically, we divided a 24-h day into four 6-h time slots—morning (*M*, 06:00–12:00), afternoon (*A*, 12:00–18:00), evening (*E*, 18:00–24:00), and night (*N*, 24:00–6:00) —for the period from 2010/07/01 to 07/31. Fig. 8 shows geographic regularities for a geographic area by means of boxplots. Specifically, we consider a median value of each boxplot only and define the value as geographic regularity. In the present experiment, we estimated 1,200 geographical regularities of *#Clusters* = 300 for every four time slots.



Dataset
(#Tweets, #Crowd, and #MovCrowd)

Geographic regularity of #Tweets

Geographic regularity of #Crowd

Geographic regularity of #MovCrowd

Fig. 8. Geographic regularity for urban areas

4.3 Characterizing Urban Areas

Our method characterizes geographic areas with respect to geographic regularities estimated by crowd activity patterns. We focused on relative change patterns of geographical regularities for clusters in terms of *#Tweets* (gr_T), *#Crowd* (gr_C), and *#MovCrowd* (gr_{MC}) between the fixed time slots. We obtained 300 change patterns from geographical regularities for the same number of clusters.

In order to find common patterns, we applied a Frequent Itemset Mining algorithm [18], statistically constructing it to full size: 9 items. In this experiment, we extracted 8 types of common change patterns whose occurrence ratios were over a threshold (here, set at 2%), as shown in Table 1. This covered 86.3% of all 300 urban areas. We finally extracted 4 types of common patterns based on the partial similarity of 8 change patterns. On the basis of these common frequent patterns, we clustered urban areas and called them “bedroom towns,” “office towns,” “nightlife towns,” and “multifunctional towns,” as shown in Table 2. We defined the characteristics based on change patterns of crowd activities as follows:

Table 1. Relative change patterns extraction based on the Frequent Itemset Mining algorithm

pattern	#Tweets			#Crowd			#MovCrowd			occurrence ratio
	p_1-p_0	p_2-p_1	p_3-p_2	p_1-p_0	p_2-p_1	p_3-p_2	p_1-p_0	p_2-p_1	p_3-p_2	
0	+	+	−	+	0	−	+	−	−	2.7
1	+	+	−	+	+	−	+	−	−	20
2	+	−	−	+	−	−	+	−	−	3
3	+	+	−	+	+	−	0	+	−	3.3
4	+	+	−	+	+	−	−	+	−	3.7
5	+	+	−	+	+	−	+	+	−	38
6	+	+	−	+	−	−	+	−	−	6.3
7	+	+	−	+	+	−	+	0	−	9.3

Table 2. Common change patterns used for clustering urban areas

pattern	#Tweets			#Crowd			#MovCrowd			occurrence ratio
	p_1-p_0	p_2-p_1	p_3-p_2	p_1-p_0	p_2-p_1	p_3-p_2	p_1-p_0	p_2-p_1	p_3-p_2	
0	+	+	−	+	0	−	+	−	−	2.7
1					+					20

(a) Bedroom town

pattern	#Tweets			#Crowd			#MovCrowd			occurrence ratio
	p_1-p_0	p_2-p_1	p_3-p_2	p_1-p_0	p_2-p_1	p_3-p_2	p_1-p_0	p_2-p_1	p_3-p_2	
2	+	−	−	+	−	−	+	−	−	3
6		+								6.3

(b) Office town

pattern	#Tweets			#Crowd			#MovCrowd			occurrence ratio
	p_1-p_0	p_2-p_1	p_3-p_2	p_1-p_0	p_2-p_1	p_3-p_2	p_1-p_0	p_2-p_1	p_3-p_2	
3	+	+	−	+	+	−	0	+	−	3.3
4							−			3.7

(c) Nightlife town

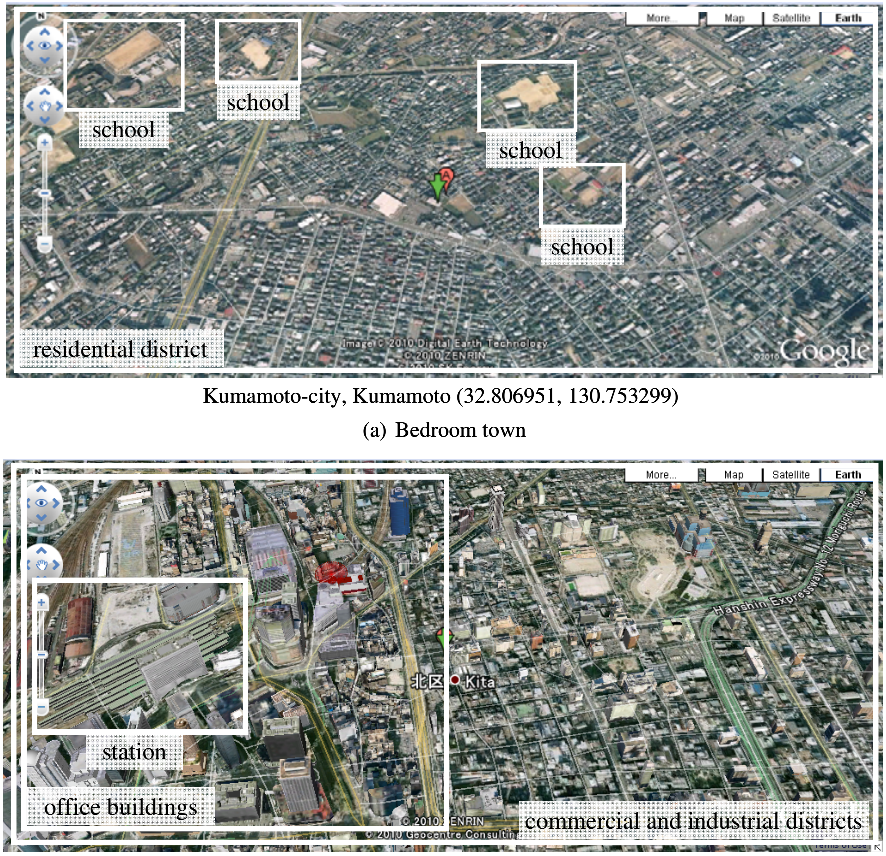
pattern	#Tweets			#Crowd			#MovCrowd			occurrence ratio
	p_1-p_0	p_2-p_1	p_3-p_2	p_1-p_0	p_2-p_1	p_3-p_2	p_1-p_0	p_2-p_1	p_3-p_2	
5	+	+	−	+	+	−	+	+	−	38
7								0		9.3

(d) Multifunctional town

(a) **Bedroom towns:** As shown in Table 2 (a), both the number of tweets and the number of crowds continue to increase in the afternoon (12:00–18:00) and evening (18:00–24:00). On the other hand, the number of moving crowds increases in the afternoon (12:00–18:00), but decreases in the evening (18:00–24:00). In short, more people exist in and come to the areas in the evening, but they do not move actively there; the activity pattern may be caused by crowds returning home after work or school. Therefore, we called the areas where active crowds aggregate in the evening and then calm down at night "bedroom towns."

(b) **Office towns:** As shown in Table 2 (b), the number of tweets, the number of crowds, and the number of moving crowds mostly increase in the afternoon (12:00–18:00), and decrease in the evening (18:00–24:00). In other words, more people exist in and come to the areas and move actively there in the afternoon, and then leave in the evening; the activity pattern can be regarded as that of crowds coming to work. Therefore, we called the areas where active crowds gather in the afternoon and disperse at night “office towns.”

(c) **Nightlife towns:** As shown in Table 2 (c), both the number of tweets and the number of crowds continue to increase in the afternoon (12:00–18:00) and evening (18:00–24:00). On the other hand, the number of moving crowds decreases in the afternoon (12:00–18:00), but increases in the evening (18:00–24:00). In short, more people exist in and come to these areas in the evening; however, they do not move actively there before the evening and then become active in the evening. The activity pattern consists of crowds who come in the evening and originally staying there for the newcomer. Therefore, we call the areas where crowds calm down in the afternoon and active crowds aggregate in the evening “nightlife towns.”


school
school
school
school
residential district
More...
Map
Satellite
Earth
Image © 2010 Digital Earth Technology
© 2010 ZENRIN
©2010 Google
station
office buildings
北区
Kita
Hanshin Expressway No. 12 Moriguchi Route
© 2010 ZENRIN
© 2010 Geocentre Consulting
commercial and industrial districts

Kumamoto-city, Kumamoto (32.806951, 130.753299)

(a) Bedroom town

Kita-ward, Osaka-city, Osaka (34.701973, 135.502744)

(b) Office town

Fig. 9. Comparing urban areas with Google Earth aerial photographs

(d) **Multifunctional towns:** As shown in Table 2 (d), the number of tweets, the number of crowds, and the number of moving crowds continue to increase in the afternoon (12:00–18:00) and evening (18:00–24:00), and then decrease in the night (24:00–06:00). In other words, more people exist in and come to the areas, move actively there until the evening, and then leave there at night; the activity pattern consists of crowds who have various purposes. Therefore, we call the areas where active crowds aggregate almost all day long, and that provide various functions and roles to support the lifestyles of multiple crowds "multifunctional towns."

To confirm the correctness of the labeling above, we looked into the regions of each cluster by means of Google Earth[3] aerial photographs. Interestingly, as depicted in Fig. 9, for "bedroom town" and "office town," we were definitely able to determine the characteristics of the crowds in those regions. In the case of "bedroom towns," the regions are characterized as "residential districts," and include some schools. Those areas that we classified as "office towns" have many high towers, indicating "commercial and industrial districts."

5 Conclusion

In this paper, we proposed a novel method for characterization of urban areas by extracting common patterns of crowd activities on Twitter. In this method, geographic regularities for urban areas were measured based on the crowd behavior on Twitter, We extracted change patterns of geographic regularities and classified urban areas that are clustered by similarities in change patterns as bedroom towns, office towns, nightlife towns, and multifunctional towns. Furthermore, we conducted an experiment using actual geo-tagged messages gathered from massive crowds all over Japan via Twitter. In order to evaluate whether the characterizations were accurate, we compared our results with the appearance of components in target urban areas. Consequently, we confirmed that crowd activities determined via Twitter can reflect and characterize living spaces in urban areas. In addition, our method would help analysts characterize geographic areas from various view points by enabling them to adjust diverse granularities of time period and region size respectively. In future work, we will explore further complex urban phenomena that can be derived from geo-social databases, using location-based social network sites.

Acknowledgments. This research was supported in part by a Grant-in-Aid for Scientific Research (B)(2) 20300039 from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

References

1. Fujisaka, T., Lee, R., Sumiya, K.: Detection of Unusually Crowded Places through Micro-Blogging Sites. In: Proc. of the 6th International Symposium on Web and Mobile Information Services (WAMIS 2010), pp. 467–472 (2010)
2. Fujisaka, T., Lee, R., Sumiya, K.: Discovery of User Behavior Patterns from Geo-tagged Micro-blogs. In: Proc. of the 4th International Conference on Ubiquitous Information Management and Communication (ICUIMC 2010), pp. 246–255 (2010)

[3] The Google earth: `http://www.google.com/earth/index.html`

3. Iwaki, Y., Jatowt, A., Tanaka, K.: Supporting finding read-valuable articles in microblogs. In: Proc. of the First Forum on Data Engineering and Information Management, A6-6 (2009) (in Japanese)
4. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: understanding micro-blogging usage and communities. In: Proc. of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, pp. 56–65 (2007)
5. Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y.: An efficient k-means clustering algorithm: analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelignece 24(7) (July 2002)
6. Krishnamurthy, B., Gill, P., Arlitt, M.: A few chirps about twitter. In: Proc. of the First Workshop on Online Social Networks (WOSN 2008), pp. 19–24 (2008)
7. Kurashima, T., Tezuka, T., Tanaka, K.: Blog Map of Experiences: Extraction and Geographical Mapping of Visitor Experiences from Urban Blogs. IPSJ SIG Technical Report, pp. 45–53 (2005)
8. Lee, R., Sumiya, K.: Measuring Geographical Regularities of Crowd Behaviors for Twitter-based Geo-social Event Detection. In: Proc. of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks (LBSN 2010), pp. 1–10 (2010)
9. Lee, R., Wakamiya, S., Sumiya, K.: Discovery of Unusual Regional Social Activities using Geo-taggged Microblogs. The World Wide Web Journal, Special Issue on Mobile Services on the Web (2011) (to appear)
10. Lynch, K.: The Image of the City (1960)
11. Mcgill, R., Tukey, J.W., Larsen, W.A.: Variations of Box Plots. The American Statistician 32(1), 12–16 (1978)
12. Moriya, K., Sasaki, S., Kiyoki, Y.: A Dynamic Creation Method of Environmental Situation Maps Using Text Data of Regional Information. DEIM Forum 2009 B1-6 (2009) (in Japanese)
13. Pew Internet, `http://pewinternet.org/~/media//Files/Reports/2010/PewInternet-OlderAdultsandSocialMedia.pdf`
14. Tezuka, T., Lee, R., Takakura, H., Kambayashi, Y.: Integrated Model for a Region-Specific Search Systems and Its Implementation. In: Proc. of 2003 IRC International Conference on Internet Information Retrieval, pp. 243–248 (2003)
15. Twitter, `http://twitter.com/`
16. Twitter open API, `http://apiwiki.twitter.com/Twitter-Search-API-Method%3A-search`
17. Vieira, M.R., Frias-Martinez, V., Oliver, N., Frías-Martínez, E.: Characterizing Dense Urban Areas from Mobile Phone-Call Data: Discovery and Social Dynamics. In: Proc. of the 2010 IEEE Second International Conference on Social Computing (SocialCom), pp. 241–248 (2010)
18. Yu, J.X., Li, Z., Liu, G.: A data mining proxy approach for efficient frequent itemset mining. The International Journal on Very Large Data Bases Archive 17(4) (July 2008)
19. Zhao, D., Rosson, M.B.: How and why people Twitter: the role that micro-blogging plays in informal communication at work. In: Proc. of the ACM 2009 International Conference on Supporting Group Work, pp. 243–252 (2009)

On the (Limited) Difference between Feature and Geometric Semantic Similarity Models

Ola Ahlqvist

Department of Geography,
The Ohio State University,
154 N Oval Mall, Columbus, OH 43210,
U.S.A.
ahlqvist.1@osu.edu

Abstract. Semantic similarity assessment is central to many geographic information analysis tasks. A reader of the geographic information science literature on semantic similarity assessment processes could easily get the impression that two of the most common approaches, the feature model and the geometric model, are incompatible and radically different. Through a review of literature I seek to elaborate on and clarify that these two approaches are in fact compatible, and I finish with a brief discussion of the handling of uncertain and missing values in these representations.

1 Introduction

Concepts are a core element in theories about how we understand and reason about of the world. Dating back as far as Aristotle, philosophy and science have primarily viewed concepts according to the "classical" view [1]; concepts are mentally represented as summary definitions where every object is either part of a category or not and all members of a concept are equally good examples of it. These and other tenets of the classical view underlie many common knowledge representation theories and logics; from Peirce's five semantic primitives; existence, coreference, relation, conjunction, negation, to first-order logic languages such SQL, and OWL [2]. Hence, to this date, the classical view of concepts provides a foundation for many important developments and technologies that provide a foundation for the semantic web vision [3].

Nevertheless, the classical view, as a theory for how humans reason, is today largely rejected after some pioneering work by Rosch [4] and others, who identified both theoretical and empirical problems with that view. One major issue with the classical view is the semantically imprecise and vague notions that are so pervasive in geography [5-8]. Although the geographic information sciences have acknowledged and worked on these semantic issues for more than a decade now [c.f. 9], most current software and information infrastructures are still very much based on the "classical" thinking, including much of more recent Geospatial web solutions [c.f. 10, 11].

The general concern from cognitive science has been articulated as; "The gradation of properties in the world means that our smallish number of categories will never map perfectly onto all objects: The distinction between member and nonmembers will

C. Claramunt, S. Levashkin, and M. Bertolotto (Eds.): GeoS 2011, LNCS 6631, pp. 124–132, 2011.

always be difficult to draw or will even be arbitrary in some cases […] if the world consists of shadings and gradations and a rich mixture of different kinds of properties, then a limited number of concepts would almost have to be fuzzy." [12]. Not surprisingly then, notions of gradedness and judgment of concept similarity is central to the three main kinds of theories that have replaced the classical view on concepts; prototype, exemplar, and the knowledge approach. All three predict that categories will have gradations of typicality and that there will be borderline cases because of that. This is a big step away from the theories that originally guided how geographic information systems were designed to model real world concepts and objects in a spatial database.

Recently, Schwering [13] reviewed five different approaches to represent and measure semantic gradation; geometric, feature, network, alignment, and transformational models. Of the five approaches, the geometric and feature based models have received significant interest in the GIScience literature, and they are also well suited to conform to both exemplar and prototype theories identified above. Some example formalizations based on the geometric and feature based frameworks can be found in [14-19].

While both models use a collection of characteristics, a geometric model defines a characteristic as a value along some dimension and the feature model is simply a list of Boolean characteristics. As an example, if a forest is characterized as an area that is tree covered, the geometric model could define a dimension called "percent tree cover" and specify the interval, say 30-100%, that characterize a forest. A feature based model could add "tree covered" to a list of characteristic features for the forest concept.

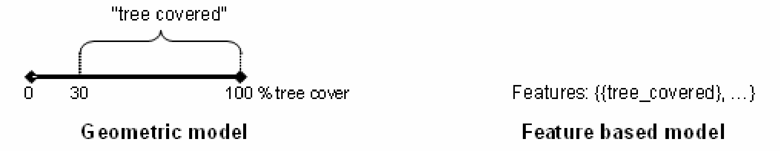


Fig. 1. The descriptive characteristic "tree covered" represented in a continuous geometric model (left) and a Boolean feature based model (right)

The evaluation of similarity would then be based on comparing tree cover for an object of interest with the concept definitions. In the geometric case a value of say 25% would be compared with the criteria 30-100% and some form of interval or other difference based metric would give an indication of the semantic similarity. We should note that the object of interest can be both a real world object with measurable attributes as well as another concept with definitional values. In the feature based model, the object of interest would be compared to the criteria of being tree covered. If that binary evaluation comes out true it would indicate semantic similarity. For most concepts similarity is evaluated based on multiple characters such that several dimensions or features are added together, and some form of weighted average is often proposed to account for different importance of the characteristics.

An interesting aspect of work on these two approaches are the efforts to eliminate some of the most commonly noted weaknesses of and differences between the geometric and feature based models. In fact, suggested modifications of both models converge in ways that make many important aspects of them largely compatible. As an example Gati and Tversky [20] followed Restle [21] to add set theoretical representations of quantitative and ordered dimensions to the feature based contrast model. They also argued that "Objects that vary along a few attributes with many ordered levels (e.g., size, elongation) are more naturally described using a dimensional language". Similarly, variations on the geometric models have sought improvements to address some of the benefits of feature based models [22]. Indeed, Smith and Medin [23] stated that "…the process that compares features should be compatible, if not virtually identical, with that which compares dimensions." Still, discussion of these two approaches in the GIScience literature seems unnecessarily entrenched in a polarized view of feature vs. geometric views, instead of acknowledging them as largely compatible frameworks. Consequently, this paper seeks to briefly elaborate on and clarify compatible capabilities of both approaches.

2 Similarity Evaluation and Scales of Measurements

A common concern about geometric similarity models is that they are interpreted to conform to some metric restrictions such as the minimality, symmetry, and triangle inequality axioms. For example, the geometric model has long been criticized for not being able to account for the asymmetric similarities that can be found in empirical psychological data [24], although that stance also has its critics [25]. Admittedly, symmetry is a property of similarity metrics on isolated geometric dimensions defined on a ratio or interval measurement scale such as the crown cover example before. However, that position can be challenged on the grounds that dimensions need not be restricted to strictly numerical dimensions. For example, in Figure 1 the geometric model is obviously more precise about what tree covered means whereas the feature model relies on a binary concept that can be interpreted very differently. We can put both models on more equal terms by using an ordinal scaled dimension in the geometric model and a collection of ordinal valued features, figure 2.

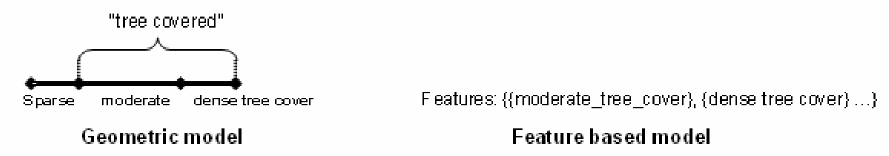


Fig. 2. The descriptive characteristic "tree covered" represented in an ordinal geometric model (left) and a feature model based on ordered sets (right)

The geometric model in figure 2 now has reduced its detail into ordinal intervals. We know intuitively that the topological structure of an ordinal dimension allow us to reason that sparse is more similar to moderate than it is to dense tree cover. More formally, we also see ordinals treated as interval values through a simple rank-order number transformation and then used by a distance based metric to calculate similarity [c.f. 26]. Although an explicit criterion according to Gärdenfors [27] is that the considered domains in a geometric model "have a metric so that we can talk about distances between points in the space”, but he also points out that “not all domains in a conceptual space are assumed to be metric.” Thus, following Velleman and Wilkinson [28], I would argue that even nominal scales in the Stevens [29] sense can be regarded as an ordered and sometimes even metric scale using said semantic similarity metrics. In fact, the whole idea about semantic similarity rests on that assumption since the categories, or nominals, are compared and evaluated in a graded and quantitative fashion. It follows then that a quality dimension can be defined using anything from a ratio to a nominal scale. However, conventional wisdom dictate that data collected over an ordinal or nominal scale can only use particular distance metrics. By recognizing and conforming to the limits of different measurement scales we can also abide to the requirement of *betweenness*, one of the major arguments against qualitative dimensions [30]. The higher detail achieved in the feature model of Figure 2 follows the idea of chained sets [20] and uses an ordered number of features, each representing a possible value, adding those features that correspond to the concept definition. Similarly, the feature based model could be extended to numerical dimensions by adding as many features necessary to represent the range of numeric values at hand. In this way, some amount of geometric, or at least topological, reasoning can be applied on the feature model as well.

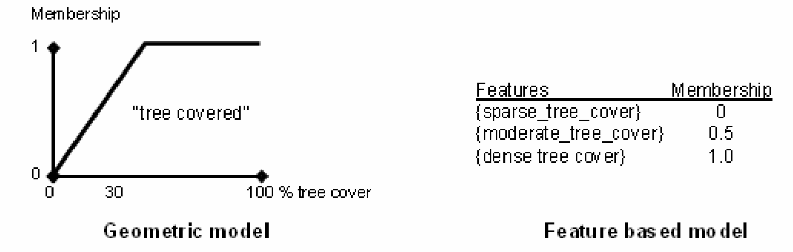
Another alternative to evaluate similarity is to follow a feature matching approach. In this we separate a dimension into distinct features/intervals and look for feature matches in the comparison. In a geometric sense this means that we look for interval overlaps in each dimension for the compared objects or concepts, and it is straightforward to implement e.g. a contrast or ratio model [24] of similarity on both ordinal, nominal or binary data using geometric reasoning [c.f. 14, 31]. While an overlap ratio can easily be computed for numerical dimensions, any ratio model (feature based or geometry based) is sensitive to the value of the denominator, e.g. the selected number of features or the total range of a dimension. In many cases there are natural bounds for a dimension, such as for the percent tree cover dimension above. Other cases where there are no set bounds there may be physical limits, such as quality dimensions for tree height or average summer temperature. Similar issues arise from choosing a particular number of features. Feature or dimension salience or prominence weights are often proposed as a way to compensate for this, but no solid theory for how this is to be done is currently available. As a side note, an important but seldom recognized aspect of an overlap ratio is that it can address category-subcategory relations [14]. For example, consider the concept ‘thicket’ could be defined as a dense growth of shrubbery or small trees. Finding that the “dense” feature, or tree cover interval, is overlapping with the “tree covered” criteria of a forest definition, we can infer that this is in fact a sub-set of the features/interval that defines a forest, indicating that 'thicket' is a subclass to 'forest'.

Returning again to the symmetry vs. asymmetry discussion, we should note that a geometric concept comparison is likely to use dimension intervals and multi-dimensional regions as the values for characteristic properties. Most difference metrics on intervals are not metric in a strict sense but actually asymmetric, and potentially able to account for cognitive asymmetries [c.f. 32]. Also note that many of the examples brought forward as examples of asymmetry (e.g. North Korea vs. Red China, Tel Aviv vs. New York) are complex concepts composed of many features or dimensions. Even if single dimensions would show symmetry, any combination of dimensions, similar to the combination of features, can compensate for varying salience or prominence of the individual dimensions or features.

3 Representation of Uncertain and Missing Values

As mentioned initially, both geographic and cognitive perspectives identify vagueness as an important element of concepts. Verguts et al. [33] also pointed out the problems associated with missing information, or null values, in many real situations. Fuzzy set theory [34] was an early candidate representation to deal with such problems but initially received some critique in how it handled prototype effects [30]. However, this early critique was based on a very simplistic notion of fuzzy set theory which has since developed to address a wide range of semantic issues; from vagueness of concept region boundaries to a full framework for quantitative fuzzy semantics [35, 36]. Several alternative representation approaches are also available including rough sets [37], rough-fuzzy sets [38], type-2 fuzzy sets [39], and supervaluation theory [40]. All of these address the idea that concepts and their attribute values are in some way vague or ambiguous. Dubois et al. [41] recently followed an information-based framework, compatible with the more modern cognitive theories mentioned above (Murphy, 2004) and argued that fuzzy set based constructs can address six kinds of vagueness, including ill-defined properties, prototype effects, agent dependence, probability distributions and information granularity. These uncertainty aspects cover most if not all of the uncertainties frequently embedded in geographic data [42].

First, using fuzzy set ideas on the geometric and feature based approaches translates into enabling individual features or (intervals of) values to be less salient to a concept definition. Again, using the forest example, we may want to define e.g. that moderate tree cover is a borderline characteristic of a forest and any object showing this character would be less similar to a ‘real’ forest. In fuzzy set parlance we could talk about salience as set membership value and a common convention is to attach a value from 0 to 1 where 1 denotes full membership and 0 denotes no membership. This allows for an explicit representation of vagueness at the attribute value level so that e.g. 30% tree cover can be regarded somewhat forest-like by associating that percentage with a membership value of 0.5 to the forest concept. As a consequence, the concept itself becomes a fuzzy set. Formally we can represent fuzzy memberships using a function on the quality dimension or a table with features and their membership values, figure 3.



Features	Membership
{sparse_tree_cover}	0
{moderate_tree_cover}	0.5
{dense tree cover}	1.0

Fig. 3. The descriptive characteristic “tree covered” represented in a continuous geometric model as a fuzzy number (left) and a fuzzy ordered feature based model (right)

Fuzzy sets like these can still be used for calculating distances over numerical dimensions using fuzzy numbers [43] or over chained sets using ranking of fuzzy sets [44]. Furthermore, Bouchon-Meunier et al. [45] provided an extensive overview of general measures for comparison of fuzzy set based descriptions of objects, including metrics that translate to the notion of overlap and class-subclass relations above.

A further aspect of using intervals like these is that descriptive metrics of these intervals can give important information based on the ranges of each property value in the concept definitions. Two useful metrics for fuzzy sets are the *core* and *support* measures. The support is defined by the interval of all non-zero membership values and the core is defined by the interval with full membership.

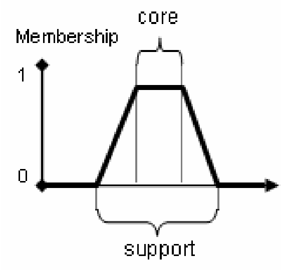


Fig. 4. The core and support measures on a fuzzy membership function

Calculating core and support measures require some measure of the range of a dimension and thus face the same issues mentioned for the ratio model before. No clear guidelines exist but as before, many cases do have natural bounds. The interpretation of these two measures has not been treated at any length in a semantic similarity context. An intuitive interpretation could be that the core represents a prototype region and the support represents the outer boundary for a concept for this particular dimension. The ratio or the difference between core and support then give an indication of the overall fuzziness of a category definition. E.g. if core/support is small, the category has a large degree of fuzziness and many borderline cases. Support and core values can also be used compare two fuzzy functions that define concepts A and B in a certain quality dimension such that a relatively larger core and support indicate a more general category.

4 Summary

The above demonstrate that both the geometric and the feature based models are capable of supporting several of the commonly suggested operations necessary to perform prototype and exemplar based semantic similarity evaluations. Still, the presence of several other representations and metrics indicate a need to look at semantic similarity from many perspectives. As an example, Rodriguez and Egenhofer [18] found it useful to combine network and feature based similarity into one compound index of similarity in their Matching-Distance Similarity Measure. Many formal ontology frameworks represent both concept relations and features to provide rich knowledge representations for a specific domain [c.f. 46] capable of calculating both hierarchical and feature based similarity metrics. Ahlqvist [47] and Ahlqvist and Gahegan [48] used an overlap and distance metric to summarize both overall similarities between concepts as well as class/subclass relationships. Computational infrastructures [c.f. 49] will need to accommodate multiple semantic similarity metrics in order to fully support multiple perspectives onto information resources.

Acknowledgements

This research was partially supported by the National Science Foundation under Grant No. BCS-0904731.

References

1. Medin, D.L.: Concepts and Conceptual Structure. Am. Psychol. 44, 1469–1481 (1989)
2. Sowa, J.F.: Knowledge representation: logical, philosophical, and computational foundations. MIT Press, Cambridge (2000)
3. Lee, T.B., Hendler, J., Lassila, O.: The semantic web. Scientific American 284, 34–43 (2001)
4. Rosch, E.: Principles of categorization. In: Rosch, E., Loyd, B.B. (eds.) Cognition and Categorization, pp. 27–48. Lawrence Erlbaum Associates, Hillsdale (1978)
5. Bennett, B.: What is a Forest? On the Vagueness of Certain Geographic Concepts. Topoi. 20, 189–201 (2001)
6. Couclelis, H.: People Manipulate Objects (but Cultivate Fields): Beyond the Raster-Vector Debate in GIS. In: Frank, A.U., Campari, I., Formentini, U. (eds.) Theories and Methods of Spatio-Temporal Reasoning in Geographic Space, pp. 65–77. Springer, Heidelberg (1992)
7. Fisher, P.: Sorites paradox and vague geographies. Fuzzy Sets Syst. 113, 7–18 (2000)
8. Fisher, P., Wood, J.: What is a Mountain? or The Englishman who went up a Boolean Geographical concept but realised it was Fuzzy. Geography 83, 247–256 (1998)
9. Salge, F.: Semantic accuracy. In: Guptill, S.C., Morisson, J.L. (eds.) Elements of Spatial Data Quality, pp. 139–151. Elsevier Science Ltd., Oxford (1995)
10. Athanasis, N., Kalabokidis, K., Vaitis, M., Soulakellis, N.: Towards a semantics-based approach in the development of geographic portals. Computers and Geosciences 35, 301–308 (2009)
11. Scharl, A., Tochtermann, K.: The Geospatial Web: How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society (Advanced Information and Knowledge Processing). Springer-Verlag New York, Inc., Secaucus (2007)
12. Murphy, G.L.: The Big Book of Concepts. MIT Press, Cambridge (2004)
13. Schwering, A.: Approaches to semantic similarity measurement for geo-spatial data: A survey. Transactions in GIS 12, 5–29 (2008)
14. Ahlqvist, O.: A Parameterized Representation of Uncertain Conceptual Spaces. Transactions in GIS 8, 493–514 (2004)
15. Feng, C.C., Flewelling, D.M.: Assessment of semantic similarity between land use/land cover classification systems. Comput., Environ. Urban Syst. 28, 229–246 (2004)
16. Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L.: Sweetening Ontologies with DOLCE. In: Gómez-Pérez, A., Benjamins, V.R. (eds.) EKAW 2002. LNCS (LNAI), vol. 2473, pp. 166–181. Springer, Heidelberg (2002)

17. Raubal, M.: Formalizing Conceptual Spaces. In: Varzi, A., Vieu, L. (eds.) Proceedings of the Third International Conference Formal Ontology in Information Systems (FOIS 2004), pp. 153–164. IOS Press, Amsterdam (2004)
18. Rodriguez, M., Egenhofer, M.: Determining semantic similarity among entity classes from different ontologies. IEEE Transactions on Knowledge and Data Engineering 15, 442–456 (2003)
19. Song, D., Bruza, P.: Towards context sensitive information inference. Journal of the American Society for Information Science and Technology 54, 321–334 (2003)
20. Gati, I., Tversky, A.: Representations of qualitative and quantitative dimensions. Journal of Experimental Psychology: Human Perception and Performance 8, 325–340 (1982)
21. Restle, F.: Psychology of Judgment and Choice. Wiley, New York (1961)
22. Johannesson, M.: Modelling asymmetric similarity with prominence. Br. J. Math. Stat. Psychol. 53, 121–139 (2000)
23. Smith, E.E., Medin, D.L.: Categories and concepts. Harvard University Press, Cambridge (1981)
24. Tversky, A.: Features of similarity. Psychol. Rev. 84, 327–352 (1977)
25. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics 19, 17–30 (1989)
26. Schwering, A., Raubal, M.: Measuring Semantic Similarity between Geospatial Conceptual Regions. In: Rodríguez, M.A., Cruz, I., Levashkin, S., Egenhofer, M.J. (eds.) GeoS 2005. LNCS, vol. 3799, pp. 90–106. Springer, Heidelberg (2005)
27. Gärdenfors, P.: Conceptual Spaces: The Geometry of Thought. MIT Press, Cambrige (2000)
28. Velleman, P.F., Wilkinson, L.: Nominal, Ordinal, Interval, and Ratio Typologies Are Misleading. The American Statistician 47, 65–72 (1993)
29. Stevens, S.S.: On the Theory of Scales of Measurement. Science 103, 677–680 (1946)
30. Osherson, D.N., Smith, E.E.: On the Adequacy of Prototype Theory as a Theory of Concepts. Cognition. International Journal of Cognitive Psychology Paris 9, 35–38 (1981)
31. Janowicz, K., Raubal, M.: Affordance-Based Similarity Measurement for Entity Types. In: Winter, S., Duckham, M., Kulik, L., Kuipers, B. (eds.) COSIT 2007. LNCS, vol. 4736, pp. 133–151. Springer, Heidelberg (2007)
32. Williams, J., Steele, N.: Difference, distance and similarity as a basis for fuzzy decision support based on prototypical decision classes. Fuzzy Sets Syst. 131, 35–46 (2002)
33. Verguts, T., Ameel, E., Storms, G.: Measures of similarity in models of categorization. Mem. Cognit. 32, 379–389 (2004)
34. Zadeh, L.A.: Fuzzy sets. Information and Control 8, 338–353 (1965)
35. Masson, M.H., Denoeux, T.: Multidimensional scaling of fuzzy dissimilarity data. Fuzzy Sets Syst. 128, 339–352 (2002)
36. Zadeh, L.A.: Quantitative fuzzy semantics. Information Sciences 3, 159–176 (1971)
37. Pawlak, Z.: Rough Sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Norwell (1992)
38. Dubois, D., Prade, H.: Rough fuzzy sets and fuzzy rough sets. International Journal of General Systems 17, 191–209 (1990)
39. Zadeh, L.A.: The Concept of a Linguistic Variable and Its Application to Approximate Reasoning. Information Sciences 8, 199–249 (1973)
40. van Fraassen, B.C.: Singular Terms, Truth-Value Gaps, and Free Logic. The Journal of Philosophy 63, 481–495 (1966)
41. Dubois, D., Esteva, F., Godo, L., Prade, H.: An information-based discussion of vagueness. Fuzzy Systems, 781–784 (2001)

42. Fisher, P.F.: Models of uncertainty in spatial data. In: Geographical Information Systems - Principles and Technical Issues, pp. 191–205. Wiley, New York (1999)
43. Tran, L., Duckstein, L.: Comparison of fuzzy numbers using a fuzzy distance measure. Fuzzy Sets Syst. 130, 331–341 (2002)
44. Bortolan, G., Degani, R.: Review of some methods for ranking fuzzy subsets. Fuzzy Sets Syst. 15, 1–20 (1985)
45. Bouchon-Meunier, B., Rifqi, M., Bothorel, S.: Towards general measures of comparison of objects. Fuzzy Sets Syst. 84, 143–153 (1996)
46. Guarino, N.: Formal ontology, conceptual analysis and knowledge representation. International Journal of Human Computer Studies 43, 625–640 (1995)
47. Ahlqvist, O.: Using semantic similarity metrics to uncover category and land cover change. In: Rodríguez, M.A., Cruz, I., Levashkin, S., Egenhofer, M.J. (eds.) GeoS 2005. LNCS, vol. 3799, pp. 107–119. Springer, Heidelberg (2005)
48. Ahlqvist, O., Gahegan, M.: Probing the relationship between classification error and class similarity. Photogramm. Eng. Remote Sensing 71, 1365–1373 (2005)
49. Gahegan, M., Pike, W.: A situated Knowledge Representation of Geographical Information. Transactions in GIS 10, 727–749 (2006)

A Facet-Based Methodology for Geo-Spatial Modeling

Biswanath Dutta, Fausto Giunchiglia, and Vincenzo Maltese

DISI - Università di Trento, Trento, Italy

Abstract. Space, together with time, is one of the two fundamental dimensions of the universe of knowledge. Geo-spatial ontologies are essential for our shared understanding of the physical universe and to achieve semantic interoperability between people and between software agents. In this paper we propose a methodology and a minimal set of guiding principles, mainly inspired by the faceted approach, to produce high quality ontologies in terms of robustness, extensibility, reusability, compactness and flexibility. We demonstrate - with step by steps examples - that by applying the methodology and those principles we can model the space domain and produce a high quality facet-based large scale geo-spatial ontology comprising entities, entity classes, spatial relations and attributes.

Keywords: space domain, methodology, principle, theory, domain ontology, geo-spatial ontology.

1 Introduction

Space and time are the two fundamental dimensions of the universe of knowledge [12, 3]. The notion of space is essential to understand the physical universe. We consider space as is in accordance with what people commonly understand by this term, which includes the surface of the earth, the space inside it and the space outside it. It comprises the usual geographical concepts, often known as features, like land formations (continents, islands, countries), water formations (oceans, seas, streams) and physiographical concepts (desert, prairie, mountain). It also comprises the areas occupied by a population cluster (city, town, village) and buildings or other man-made structures (school, bank, mine).

Spatial (geo-spatial) and temporal ontologies, because representing a shared understanding of a domain [10], are essential to achieve semantic interoperability between people and between applications. Equally important, the definition of entity types and corresponding properties has become a central issue in data exchange standards where a considerable part of the semantics of data may be carried by the categories that entities are assigned to [20]. As a matter of fact, current standards - for instance the specifications provided for the geographical domain by the Open Geospatial Consortium (OGC)[1] - do not represent an effective solution to the interoperability problem. In fact, they only aim at syntactic agreement [11] by fixing the standard terms and not allowing for variations on the terminology to be used.

[1] `http://www.opengeospatial.org/`

C. Claramunt, S. Levashkin, and M. Bertolotto (Eds.): GeoS 2011, LNCS 6631, pp. 133–150, 2011.

Several frameworks have been proposed to build and maintain geo-spatial ontologies [13, 14, 15, 21], and we also recently proposed our multilingual ontology, called GeoWordNet, that overcomes their qualitative and quantitative limitations (as extensively described in [2]). However, to the best of our knowledge no systematic ways, i.e. based on a well founded methodology and guiding principles, for building geo-spatial ontologies have been proposed so far.

Our main contribution is a methodology and a minimal set of guiding principles aimed at modelling the spatial domain and at building the corresponding background knowledge taking into account the classes, the entities, their relations and properties. As explained across this paper, the domain knowledge is organized following the well founded *faceted approach* [3], borrowed from library and information science. Note that the methodology and the guiding principles we propose are not only applicable to the spatial domain, but across domains. In this approach, the analysis of the domain allows the identification of the basic classes of real world objects. They are arranged, per *genus et differentia* (i.e. by looking at their commonalities and their differences), to construct specific ontologies called *facets*, each of them codifying a different aspect of the domain at hand. This allows being much more rigorous in the definition of the domain and its parts, in its maintenance and use [1]. The intended use of this background knowledge is manifold. Identifying the domain specific terminology and corresponding entity names allows using them to annotate, index and search geographical resources as well as for word sense disambiguation.

The rest of the paper is organized as follows. In Section 2 we illustrate our methodology and the guiding principles we propose to model a domain. In Section 3, with some step by step examples, we highlight some of the issues we faced in building the space domain. In Section 4 we describe how we further organize the elements of the domain into three main categories: entity classes, relations and attributes. Section 5 provides some statistics about the space domain, as we modelled it so far. Section 6 concludes the paper and provides our future research directions.

2 The Methodology

Our methodology is mainly inspired by the *faceted approach* proposed by the Indian librarian Ranganathan [3] at the beginning of the last century. In this approach, the domain under examination is decomposed into its basic constituents, each of them denoting a different *aspect of meaning*. Each of these components is called a *facet*. More precisely, a facet is a hierarchy of homogeneous terms, where each term in the hierarchy denotes a primitive atomic concept, i.e. a primitive class of real world objects. In the next two sections we describe the main steps in the creation of the set of facets for a given domain and the guiding principles to be used.

2.1 Steps in the Process

The building process is organized into subsequent phases as follows:

- ***Step 1: Identification of the atomic concepts***. It consists in collecting the terms representing the relevant (according to the purpose) real world entities of the domain at hand. Each term denotes a class of objects. In general, this is mainly done by interviewing domain experts and by reading available literature on that particular domain including *inter-alia* indexes, abstracts, glossaries, reference works. Analysis of query logs, when available, can be extremely valuable to determine user's interests. Each term is analyzed and disambiguated into an atomic concept. This can be approximated by associating a natural language definition to each of them. For instance, *river* can be defined as *"a large natural stream of water (larger than a brook)"* and represents the set of all real world rivers.

- ***Step 2: Analysis.*** The atomic concepts are analyzed per *genus et differentia*, i.e. in order to identify their commonalities and their differences. The main goal is to identify as many distinguishing properties - called *characteristics* - as possible of the real world entities represented by the concepts. This allows being as fine grained as wanted in differentiating among the concepts. For instance, for the concept *river* we can identify the following characteristics:

  - a body of water
  - a flowing body of water
  - no fixed boundary
  - confined within a bed and stream banks
  - larger than a brook

- ***Step 3: Synthesis.*** The synthesis aims at arranging the atomic concepts into *facets* by characteristic. At each level of the hierarchy - each of them representing a different level of abstraction - similar concepts are grouped by a common characteristic. Concepts sharing the same characteristic form an *array* of homogeneous concepts. Concepts in each array can be further organized into sub-groups (or subfacets) generating a new level in the hierarchy. Children are connected to their parent through a *genus-species* (is-a) or *whole-part* (part-of) relation. For instance, due to their commonalities we could place in the same array the concept *river* and the concept *brook.*

- ***Step 4: Standardization.*** Each atomic concept can be potentially denoted with different words. When more than one candidate is available, a standard (or preferred) term should be selected among the synonyms. This is usually done by identifying the term which is most commonly used in the domain and which minimizes the ambiguity. This is similar to the WordNet[2] approach where terms are ranked in the synset and the first one is the preferred. For instance, the concept *pharynx*, defined as *"the passage to the stomach and lungs; in the front part of the neck below the chin and above the collarbone"*, can be denoted also with *throat*. However, *pharynx* is the one most commonly used by subject specialists in the medicine domain.

- ***Step 5: Ordering.*** Concepts in each array are ordered. There are many criteria one may follow, e.g., by chronological order, by spatial order, by increasing and decreasing quantity (for instance by size), by increasing complexity, by canonical order, by literary warrant and by alphabetical order. The sequencing criteria should be based upon the purpose, scope and subject of the classification system.

[2] `http://wordnet.princeton.edu/`

For example, since the purpose of the medicine domain is to prevent and cure the diseases that can affect the human body, the facets in the domain can be, in order: body and its organs, diseases and treatments.

Following the steps above leads to the creation of a set of facets. They constitute a *faceted representation scheme* for the domain. A faceted representation scheme codifies the basic building blocks that can be used - at indexing, classification and searching time - to construct complex labels, called *subjects*. This is what in library science is called post-coordination, in contrast to pre-coordination, as it is pursued by classical enumerative approaches, where a totally new concept is added to the scheme each time a new subject has to be included. Pre-coordination clearly leads to an exponential explosion in the number of subjects, while in the faceted approach they are instead created by composing the atomic concepts from the facets. A faceted representation scheme corresponds to what in our previous work we call the *background knowledge* [4, 5], i.e. the a-priori knowledge which must exist to make semantics effective. Each facet corresponds to what in logics is called *logical theory* [23, 24] and to what in computer science is called *ontology*, or more precisely *lightweight ontology* [6].

2.2 Guiding Principles

In this section we propose a minimal set of guiding principles for building facet-based domain ontologies. These principles are derived from the canons postulated by Ranganathan in his work on prolegomena to library classification [3]. Originally he proposed a huge amount of canons and principles, with a lot of redundancy and complexities. However, instead of going into the technicalities of all of them, here we rather prefer to summarize them into a minimal set of basic principles to be followed:

1. ***Relevance.*** The selection of the characteristics to form the facets in the scheme from the atomic concepts should reflect the purpose, scope and subject of the classification system. For example, while the characteristic *by grade* looks appropriate to classify the universe of boys and girls in the context of the education domain, for sure it is not suitable to classify the universe of cows. In the latter case *by breed* would be more realistic. It is worthwhile also noting that the selection of characteristics should be done carefully, as they cannot be changed unless there is a change in the purpose, scope and subject of the classification system.

2. ***Ascertainability.*** Characteristics must be definite and ascertainable. For example, the characteristic *flowing body of water* for rivers can be ascertained easily from the scientific literature and from the geo-scientists.

3. ***Permanence.*** Each characteristic should reflect a permanent quality of an entity. For example, a spring (*a natural flow of ground water*) is always a flowing body of water, thus the facet *flowing body of water* represents a permanent characteristic of spring.

4. ***Exhaustiveness.*** Classes in an array of classes and the sub-classes in an array of sub-classes should be totally exhaustive w.r.t. their respective common immediate universe. For example, to classify the universe of people *by gender*, we

need both *male* and *female*. If we miss any of these two, the classification becomes incomplete. Note that we are not pretending to achieve such exhaustiveness in advance. The identification of the classes is based on the known real world entities. It is always possible to extend the classification in the future.

5. ***Exclusiveness.*** All the characteristics used to classify an entity must be *mutually exclusive*, i.e. no two facets can overlap in content. For example, the universe of people cannot be classified by both the characteristics *age* and *date of birth*, as they produce the same divisions.

6. ***Context.*** The denotation of a term is determined by its position in a classification system. This principle is particularly helpful for distinguishing the homographs, i.e. same term but totally different meanings. See for instance how we solve the ambiguity of the term *bank* in Section 3.4.

7. ***Currency.*** The terms denoting the classes and sub-classes should be those of current usage in the subject field. For example, in the context of transportation systems, *metro station* is more commonly used than *subway station*.

8. ***Reticence.*** The terms used to denote the classes and sub-classes should not reflect any bias or prejudice (e.g. of gender, cultural, religious), or express any personal opinion of the person who develops the classification system. For example, it is not appropriate to use terms like *minor author* or *black man*.

9. ***Ordering.*** The order should reflect the purpose, scope and subject of the classification system. Also, the ordering of facets should be consistent and should not be changed unless there is a change in the purpose, scope or subject of the classification system. Note that ordering carries semantics as it provides implicit relations between coordinate (siblings) terms.

Following the principles guarantees the creation of high quality domain ontologies in terms of robustness, extensibility, reusability, compactness and flexibility [3, 25, 26].

3 The Space Domain

Following the steps and the principles described in the previous section, we created a faceted representation scheme for the space domain.

3.1 Identification of the Atomic Concepts

Similarly to any other domain, our first step was to collect the terms and to identify the corresponding concepts representing real world geographical entities. For instance, the term *lake* corresponds to the concept *"a body of (usually fresh) water surrounded by land"* (as it is defined in WordNet) and represents the set of all real world lakes. To collect such terms we mainly used GeoNames[3] and WordNet (version 2.1).

[3] `http://www.geonames.org`

We also occasionally used the Getty Thesaurus of Geographical Names (TGN)[4] and referred to domain specific scientific literature to solve ambiguous cases.

- ***GeoNames*** is one of the most famous geo-spatial databases. It includes over 8 millions of place names in multiple languages. It also provides corresponding properties such as latitude, longitude, altitude and population. At top level, the places are categorised into 9 feature classes, further divided into 663 sub-classes.

- **WordNet** is the Princeton lexical database for the English language. WordNet groups words of different part of speech (nouns, verbs, adjectives and adverbs) into sets of cognitive synonyms, called synsets, each expressing a distinct concept. Basically, each synset groups all the words with same meaning or sense. Synsets are interlinked by means of conceptual-semantic and lexical relations. Typical semantic relations are *hypernym* (is-a) and *part meronym* (part-of). An example of lexical relation is *Participle of verb*.

- **TGN** is a structured vocabulary for place names. Similarly to GeoNames it provides around 1.1 millions of place names and 688 feature classes. It includes administrative political (e.g., cities, nations) and physical (e.g., mountains, rivers) entities. It focuses on places particularly important for the study of art and architecture.

As a preliminary step, we mapped GeoNames feature classes with WordNet synsets. From their integration we created GeoWordNet, one of the biggest multi-lingual geo-spatial ontologies currently available and therefore particularly suitable to provide semantic support for spatial applications. A large subset of GeoWordNet is available as open source[5] in plain CSV and RDF formats. This mapping allowed, among other things, identifying the main subtrees of WordNet containing synsets representing geographical classes. These are rooted in:

- **location** - a point or extent in space
- **artifact, artefact** - a man-made object taken as a whole
- **body of water, water -** the part of the earth's surface covered with water (such as a river or lake or ocean); "they invaded our territorial waters"; "they were sitting by the water's edge"
- **geological formation, formation** - the geological features of the earth
- **land, ground, soil** - material in the top layer of the surface of the earth in which plants can grow (especially with reference to its quality or use); "the land had never been plowed"; "good agricultural soil"
- **land, dry land, earth, ground, solid ground, terra firma** - the solid part of the earth's surface; "the plane turned away from the sea and moved back over land"; "the earth shook for several minutes"; "he dropped the logs on the ground"

[4] http://www.getty.edu/research/conducting_research/vocabularies/tgn

[5] http://semanticmatching.org/download.html

It is worthwhile to underline that not all the nodes in these sub-trees necessarily need to be part of the space domain. As a matter of fact, most of the descendants of *location* and *artifact* do not fall under the space domain. For instance the following:

(Descendants of location)

- **there** - a location other than here; that place; "you can take it from there"
- **somewhere** - an indefinite or unknown location; "they moved to somewhere in Spain"
- **seat** - the location (metaphorically speaking) where something is based; "the brain is said to be the seat of reason"

(Descendants of artifact)

- **article** - one of a class of artifacts; "an article of clothing"
- **anachronism** - an artifact that belongs to another time
- **block** - a solid piece of something (usually having flat rectangular sides); "the pyramids were built with large stone blocks"

3.2 Analysis

The purpose of the analysis is to enlist the characteristics to be used to form the facets. In other words they are used to form the different levels of abstraction of the conceptual categories. Real world geographical entities were analyzed using their topological, geometric or geographical characteristics. We tried to be exhaustive in their determination. This leaves open the possibility to form a huge number of very fine grained groups of atomic concepts.

In order to illustrate the analysis process, consider the following list of real world geographical entities and their corresponding glosses.

- **Mountain** - a land mass that projects well above its surroundings; higher than a hill
- **Hill** - a local and well-defined elevation of the land; "they loved to roam the hills of West Virginia"
- **Stream** - a natural body of running water flowing on or under the earth
- **River** - a large natural stream of water (larger than a brook); "the river was navigable for 50 miles"

Following the principles provided in the previous section, it is not difficult to derive the following characteristics:

- **Mountain characteristics:**
  - the well defined elevated land
  - formed by the geological formation (where geological formation is a natural phenomenon)
  - altitude in general >500m

- **Hill characteristics:**
  - the well defined elevated land
  - formed by the geological formation, where geological formation is a natural phenomenon
  - altitude in general <500m
- **Stream characteristics:**
  - a body of water
  - a flowing body of water
  - no fixed boundary
  - confined within a bed and stream banks
- **River characteristics:**
  - a body of water
  - a flowing body of water
  - no fixed boundary
  - confined within a bed and stream banks
  - larger than a brook

3.3 Synthesis

Consider the list of characteristics selected with the analysis. The first characteristic of each of the concepts above clearly suggests the distinction between two basic categories, the first consisting of the concepts *mountain* and *hill* and the second consisting of the concepts *stream* and *river*. Based upon those characteristics, two facets can be formed. They can be named as *natural elevation* and *flowing body of water* respectively. A further analysis of the characteristics suggested the creation of the more generic facets *landform* and *body of water* respectively.

The concepts *mountain* and *hill* can be further differentiated *by size*. Note that, according to the guiding principles, size is a good distinguishing characteristic for the space domain. In fact, it can be considered (almost) permanent in nature. Note that this is not true in general. For instance, it is not appropriate to distinguish animals by size because in this respect size is transitional in nature, i.e. their size rapidly changes over time. This is an example of what Aristotle called *accidental predicates* [16].

Note that *river* is a natural stream, and therefore a special kind of *stream*. In particular, this means that all the properties of stream are inherited by river (but not the vice versa). This is reflected in the facet by putting *river* under *stream*.

Based upon the observations above we can build the following classification scheme with two facets, *body of water* and *landform*:

Body of water
- Flowing body of water
  - Stream
    - River

Landform
- Natural elevation
  - Mountain
  - Hill

An important property of facets is that they are *hospitable* (the interested reader can refer to [1] for the list of the most important properties of facets), i.e. they can be easily extended to accommodate additional concepts as needed. Assume for instance that the new concept *lake* (*a body of (usually fresh) water surrounded by land*) is identified. By analyzing it, we can derive the following characteristics:

- **Lake characteristics:**
  - a body of fresh water
  - fixed geographical boundary
  - a stagnant body of water

Going through the characteristics above, it should be easy to understand that *lake* cannot be put under the *flowing body of water*, even though it is a *body of water*. This implies that our classification is not good enough to classify all kinds-of body of water, i.e. it is not exhaustive (principle of exhaustiveness). In order to include lakes, we need to extend the body of water facet with *stagnant body of water* in the same array of *flowing body of water*. This solves our problem.

In order to understand the importance of the principle of exclusiveness, assume to create in our classification the sub-classes *inland body of water*, *marine body of water*, *flowing body of water*, and *stagnant body of water* in the same array level under the main class *body of water*. Such categorization brings to confusion. In fact, lake can be now classified as both *inland body of water* and *stagnant body of water*. To avoid this confusion, the principle of exclusiveness plays an important role. According to this principle, all the characteristics used to classify an entity must be *mutually exclusive*. So, we should not include all those four sub-classes in the same array.

Similarly to lakes, we can extend the *natural elevation* facet in order to accommodate the concept *valley* (*a long depression in the surface of the land that usually contains a river*). Valley is a natural depression. So, in order to assign a place for *valley* inside this scheme, we have to create another sub-facet, namely, *natural depression*.

Consider that valleys are seen in both the oceanic areas (called *oceanic valley*) and continental areas (called *valley*). There is in general symmetry of real world entities in the continental and oceanic areas. For most of the continental entity classes there is a corresponding oceanic entity class with similar features but different name. So, in order to correctly classify the entities based upon the characteristic of their location, i.e. oceanic or continental, we should create the sub-facets oceanic and continental under the natural elevation and natural depression respectively as shown below. These additional facets make the classification of *landforms* exhaustive.

Body of water
- Flowing body of water
  - Stream
    - Brook
    - River
- Stagnant body of water
  - Pond
  - Lake

Landform
- Natural depression
  - Oceanic depression
    - Oceanic valley
    - Oceanic trough
  - Continental depression
    - Trough
    - Valley

- Natural elevation
  - Oceanic elevation
    - Seamount
    - Submarine hill
  - Continental elevation
    - Hill
    - Mountain

By applying more and more characteristics of division, the extension of the concepts decreases and the intension increases. For example, there are fewer kinds-of *lake* than kinds-of *stagnant body of water*. See the appendix for a complete example.

3.4 Standardization

For each concept a standard term was selected while all the others are still kept as synonyms. This allows variations supporting semantic interoperability between systems using different terminology. For the concepts extracted from WordNet, we followed the order of the words in the corresponding synsets. For the concepts extracted from GeoNames we either kept the original terms - if found appropriate - or we changed them based on the study of some scientific publications. For instance, we changed *mountains* (from the feature class T, including land formations) into *mountain range* (as from Geology terminology), and *hill* (from the feature class U, including undersea entities) into *submarine hill* (as from Oceanography terminology). Some other examples and the criteria we used can be found in [2]. For the remaining concepts we used standard vocabularies.

In general it is good practice to avoid choosing the same standard term to denote two totally different concepts within a domain. However, in one case - for the word *bank* - we had to allow an exception:

- **bank** - sloping land (especially the slope beside a body of water)) *"they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents"*
- **bank** - a building in which the business of banking transacted; *"the bank is on the corner of Nassau and Witherspoon"*

In these extreme cases, it is the context that disambiguates their meaning (principle of context). The two meanings of bank were disambiguated as follows:

- **Landform >** Natural elevation > Continental elevation > Slope > Bank
- **Facility >** Business establishment > Bank

3.5 Ordering

Given our purpose and scope, we ordered the classes based upon the *decreasing quantity* of the entities instantiating the class. Within each chain of concepts, from the root to the leaves, we followed the same ordering preference. However, it is not always possible or appropriate to establish this order, especially when the classes do not share any characteristic. For example, we could not establish an order between *body of water* and *landform*. In such cases we preferred the *canonical order*, i.e. the order traditionally followed in Library Science. The final result, after ordering, was as follows:

Landform

- Natural elevation
  - Continental elevation
    - Mountain
    - Hill
  - Oceanic elevation
    - Seamount
    - Submarine hill
- Natural depression
  - Continental depression
    - Valley
    - Trough
  - Oceanic depression
    - Oceanic valley
    - Oceanic trough

Body of water

- Flowing body of water
  - Stream
    - River
    - Brook
- Stagnant body of water
  - Lake
  - Pond

4 Elements of the Space Domain

The faceted representation scheme we created represents *classes* of real world geographical entities. To complete our model of the domain we also provide in this section the *relations* between them and their *attributes*. We consider classes, relations, and attributes as the three fundamental components, or categories, of any domain.

4.1 Entity Classes

This category contains the classes of the faceted representation scheme. It is the main means to determine what an object is. In other words, we can characterize each real world geographical entity by associating it to its entity class. The space domain consists of the following basic facets:

- **Region** - a large indefinite location on the surface of the Earth; "penguins inhabit the polar regions"
- **Administrative division** - a district defined for administrative purposes
- **Populated place** - a city, town, village, or other agglomeration of buildings where people live and work
- **Facility** - a building or any other man-made permanent structure that provides a particular service or is used for a particular industry; "the assembly plant is an enormous facility"

- **Abandoned facility** - abandoned or ruined building and other permanent man made structure which are no more functional
- **Land** - material in the top layer of the surface of the earth in which plants can grow (especially with reference to its quality or use); *"the land had never been plowed"; "good agricultural soil"*
- **Landform** - the geological features of the earth
- **Body of water** - the part of the earth's surface covered with water (such as a river or lake or ocean) "they invaded our territorial waters"; "they were sitting by the water's edge"

Each of these top-level facets is further sub-divided into several sub-facets. For example, *facility* is sub-divided into *living accommodation*, *religious facility*, *education facility*, *research facility*, *education research facility*, *medical facility*, *transportation facility*, and so on. Similarly, *body of water* is further sub-divided primarily into the two sub-facets *flowing body of water* and *stagnant body of water.* In a similar way, *landform* is further subdivided into the two sub-facets *natural elevation* and *natural depression.* At lower levels all of them are further sub-divided into sub-sub-facets and so on. For example, *natural elevation* consists of the sub-facets *continental elevation* and *oceanic elevation*, while *natural depression* consists of the sub-facets *continental depression* and *oceanic depression.*

4.2 Relations

The real world entities indeed exist in the real world and they occupy some region of space on the earth surface. It is quite natural to describe how objects are located in space in relation to other objects. Understanding spatial relations is one of the fundamental features of Geographic Information Systems (GIS). According to Egenhofer and Herring [19], spatial regions form a relational system comprising the relations between interiors, exteriors, and boundaries of two objects. Spatial relations play an important role for effective geographical knowledge discovery. Consider for instance the following queries:

- "R*etrieve all the secondary schools within 500 meters of the Dante railway station in Trento*"
- "*Find all the highways of the Trentino province adjacent to marine areas*".

Since people tend to express and understand spatial relations through natural language [8], we also expressed them accordingly. Arpinar et al. [8] suggest three major types of spatial relations: topological relations, cardinal direction and proximity relations. Egenhofer and Dupe [9] propose topological and directional relations. According to them, topological properties have a leading role in qualitative spatial reasoning. Pullar and Egenhofer in [7] group spatial relations into direction relations (e.g. north, north-east), topological relations (e.g. disjoint), comparative or ordinal relations (e.g. in, at), distance relations (e.g. far, near) and fuzzy relations (e.g. next to, close).

The spatial relations we propose can be compared to the work in [7]. However, in addition to the standard direction, topological, ordinal, distance and fuzzy relations, we extend them by including relative level (e.g. above, below), longitudinal (e.g. in front, behind), side-wise (e.g. right, left), position in relation to border or frontier (e.g. adjacent, overlap) and other similar relations. A partial list of the spatial relations we propose is reported in Table 1, organized in a faceted fashion.

Note that in addition to the spatial relations, we also consider some other kinds of relations, which can be treated as functional relations. For example, in the context of lakes, primary inflow and primary outflow are two important relations.

Table 1. Partial list of spatial relations

Direction	East South-east South South-west …
Internal spatial relation	Inside Central - Midpoint - Midplane - Concentric - Eccentric …
External spatial relation	Alongside Adjacent Near Neighbourhood ...
Position in relation to a border or frontier	Adjacent (touching) Overlap Opposite …
Longitudinal spatial relation	In front Mid-length (amidships) Behind In line Toward …
Sideways spatial relation	Right (right side) Centre-line Left Alongside Across …
Relative level	Above Below Up ...

4.3 Attributes

An attribute is an abstraction belonging to or a characteristic of an object. This is a construct through which objects or individuals can be distinguished. Attributes are therefore effective for Named Entity Recognition (NER) [18] and for efficient geographical information retrieval (GIR) [17]. For example, there are 14 locations called Rome in United States of America (USA), one in Italy (the capital city of Italy) and one in France. Using the latitude and longitude attributes stored in the background knowledge - for instance GeoWordNet - we can easily distinguish them.

Attributes are primarily *qualitative* and *quantitative* in nature. For example, we may mention depth (of a river), surface area (of a lake), length (of a highway) and altitude (of a hill). For each of these attributes, we may have both qualitative and quantitative values. We store the possible qualitative values in the background knowledge. This provides a controlled vocabulary for them. They are mostly *adjectives*. For example, for depth (of a river) the possible values are {wide, narrow}. Similarly, for altitude (of a hill) the possible values are {high, low}.

We also make use of *descriptive* attributes. They are used to describe, usually with a short natural language sentence, a specific aspect of an entity. Typical examples are the history (of a monument) or the architectural style (of a building) or any user defined tag.

5 Statistics

In this section we report some statistics about our space domain. Table 2 provides the total number of objects we identified. Note that for the relations we do not count the taxonomical *is-a* and *part-of* relations. Similarly, for the attributes we do not count the number of attribute values, but only the attribute names. As part of this work, the faceted representation scheme we developed has been aligned with GeoWordNet and it is used to classify its 6,907,417 locations. This provides a faceted infrastructure to index, browse and exploit GeoWordNet. We are further increasing this number by importing locations from other sources. For instance, with the SGC project in collaboration with the Autonomous Province of Trento in Italy, a dataset of 20,162 locations of the province has been analyzed and integrated with GeoWordNet [22]. Table 3 provides a fragment of the scheme populated with the locations from GeoWordNet.

Table 2. Statistics of the Space domain

Objects	Quantity
Entity classes	845
Relations	70
Attributes	35
Locations	6,907,417

In comparing our space domain with the existing reputed and popularly used geospatial ontologies, like GeoNames and TGN, our space domain is much richer in all its aspects. Just to provide a small glimpse, GeoNames and TGN count 663 and 688 classes respectively; while in our domain we have, at this stage, 845 classes. Our plan is in fact to further increase the coverage of our space domain, both in terms of entities, entity classes, arbitrary relations and attributes. This allows a more and more accurate annotation, disambiguation, indexing and search on geographical resources. It is worthwhile to underline that, since hospitality is one of the significant features of our representation scheme, we can extend the domain at any given point of time and at any extend of granularity as we want to be.

Table 3. A fragment of the populated scheme

Objects	Quantity
Mountain	279,573
Hill	158,072
Mountain range	19,578
Chain of hills	11,731
Submarine hills	78
Chain of submarine hills	12
Oceanic mountain	5
Oceanic mountain range	0

6 Conclusion

Starting from the observation that ontologies are fundamental to achieve semantic interoperability in a domain, and that many attempts have been already made towards building geo-spatial ontologies, we have emphasized the need to follow a systematic approach - based on a well founded methodology and guiding principles - to ensure high quality results. We have presented our methodology and guiding principles, mainly inspired by the faceted approach, used for several decades and currently in use with great success in the library and information science field. By applying the methodology we modelled the space domain as a faceted representation scheme where the main components are the entities, the entity classes, their relations and attributes. By comparing our result w.r.t. well known geographical resources, like GeoNames and TGN, we have shown how, in all its components, our coverage is much bigger and our quality (as a well established feature of the methodology followed) is much better.

As future work, we plan to further extend the coverage of our space domain, in terms of entities, entity classes, relations and attributes. This will be achieved mainly from the analysis of the WordNet concepts not considered during the first phase of our work and by importing entities from other sources.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 231126 LivingKnowledge: LivingKnowledge - Facts, Opinions and Bias in Time.

References

1. Giunchiglia, F., Dutta, B., Maltese, V.: Faceted Lightweight Ontologies. In: Borgida, A.T., Chaudhri, V.K., Giorgini, P., Yu, E.S. (eds.) Conceptual Modeling: Foundations and Applications. LNCS, vol. 5600, pp. 36–51. Springer, Heidelberg (2009)
2. Giunchiglia, F., Maltese, V., Farazi, F., Dutta, B.: GeoWordNet: A resource for geo-spatial applications. In: Aroyo, L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L., Tudorache, T. (eds.) ESWC 2010. LNCS, vol. 6088, pp. 121–136. Springer, Heidelberg (2010)
3. Ranganathan, S.R.: Prolegomena to library classification. Asia Publishing House (1967)
4. Giunchiglia, F., Shvaiko, P., Yatskevich, P.: Discovering Missing Background Knowledge in Ontology Matching. In: Proceedings of the 17th European Conference on Artificial Intelligence - ECAI 2006 (2006)
5. Giunchiglia, F., Kharkevich, U., Zaihrayeu, I.: Concept Search: Semantics Enabled Syntactic Search. In: Semantic Search 2008 Workshop (SemSearch2008) at the 5th European Semantic Web Conference, ESWC (2008)
6. Giunchiglia, F., Zaihrayeu, I.: Lightweight ontologies. In: Ozsu, M.T., Liu, L. (eds.) Encyclopedia of Database Systems. Springer, Heidelberg (2008)
7. Pullar, D., Egenhofer, M.J.: Toward formal definitions of topological relations among spatial objects. In: Proceedings of the 3rd International Symposium on Spatial Data Handling, Sydney, Australia, pp. 165–176 (1988)
8. Arpinar, I.B., Sheth, A., Ramakrishnan, C.: Geospatial ontology development and semantic analytics. In: Wilson, J.P., Fotheringham, A.S. (eds.) Handbook of Geographic Information Science. Blackwell Pub., London (2004)
9. Egenhofer, M.J., Dube, M.P.: Topological relations from metric refinements. In: ACM GIS, Seattle, WA, USA (2009)
10. Gruber, T.R.: Toward Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal of Human and Computer Studies 43(5/6), 907–928 (1995)
11. Kuhn, W.: Geospatial semantics: Why, of what, and how? In: Spaccapietra, S., Zimányi, E. (eds.) Journal on Data Semantics III. LNCS, vol. 3534, pp. 1–24. Springer, Heidelberg (2005)
12. Maltese, V., Giunchiglia, F., Denecke, K., Lewis, P., Wallner, C., Baldry, A., Madalli, D.: On the interdisciplinary foundations of diversity. In: At the first Living Web Workshop at ISWC 2009 (2009)
13. Abdelmoty, A.I., Smart, P., Jones, C.B.: Building Place Ontologies for the Semantic Web: issues and approaches. In: Proc. of the 4th ACM Workshop on GIR (2007)
14. Auer, S., Lehmann, J., Hellmann, S.: LinkedGeoData: Adding a spatial dimension to the web of data. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 731–746. Springer, Heidelberg (2009)
15. Chaves, M.S., Silva, M.J., Martins, B.: A Geographic Knowledge Base for Semantic Web Applications. In: Proc. of 20th Brazilian Symposium on Databases, SBBD (2005)
16. Smith, B., Mark, D.M.: Ontology and geographic kinds. In: Proc. of the International Symposium on Spatial Data Handling, Vancouver, Canada (1998)
17. Jones, C.B., Abdelmoty, A.I., Fu, G.: Maintaining Ontologies for Geographical Information Retrieval on the Web. In: Chung, S., Schmidt, D.C. (eds.) CoopIS 2003, DOA 2003, and ODBASE 2003. LNCS, vol. 2888, pp. 934–951. Springer, Heidelberg (2003)
18. Kalfoglou, Y., Schorlemmer, M.: Ontology mapping: the state of the art. Knowledge Engineer. Review 18(1), 1–31 (2003)

19. Egenhofer, M., Herring, J.: Categorization binary topological relationships between regions, lines, and points in geographic databases. In: Egenhofer, M., Herring, J. (eds.) A Framework for the Definition of Topological Relationships and an Approach to Spatial Reasoning within this Framework, Santa Barbara, CA (1991)
20. Mark, D.M.: Toward a theoretical framework for geographic entity types. In: Frank, A.U., Campari, I. (eds.) COSIT 1993. LNCS, vol. 716, pp. 270–283. Springer, Heidelberg (1993)
21. Duce, S.: Towards an Ontology for Reef Islands. In: Proceedings of the 3rd International Conference on GeoSpatial Semantics (2009)
22. Farazi, F., Maltese, V., Giunchiglia, F., Ivanyukovich, A.: A semantic geographical catalogue for semantic search. DISI Technical report (2010)
23. Giunchiglia, F., Villafiorita, A., Walsh, T.: Theories of Abstraction. AI Communications 10(3/4), 167–176 (1997)
24. Giunchiglia, F., Walsh, T.: Abstract Theorem Proving. In: Proceedings of the 11th International Joint Conference on Artificial Intelligence (IJCAI 1989), pp. 372–377 (1989)
25. Broughton, V.: The need for a faceted classification as the basis of all methods of information retrieval. Aslib Proceedings 58(1/2), 49–72 (2006)
26. Spiteri, L.: A Simplified Model for Facet Analysis. Journal of Information and Library Science 23, 1–30 (1998)

Appendix: The Complete Body of Water Facet

Body of water

- Ocean
- Sea
  - Bay
- Bight
- Gulf
- Inlet
  - Cove
- Flowing body of water
  - Stream
    - River
      - Lost river
    - Brook
      - Brooklet
      - Tidal brook
    - Headstream
    - Rivulet
    - Branch
      - Anabranch
      - Billabong
      - Distributory
      - Tributory
    - Canalized stream
    - Tidal stream
    - Intermittent stream
  - Channel
    - Watercourse
      - Abandoned watercourse
    - Navigation channel
    - Reach
    - Marine channel
    - Lake channel
    - Cutoff
  - Overfalls
  - Current
    - Whirlpool
  - Section of stream
    - Headwaters
    - Confluence
    - Stream mouth
      - Estuary
    - Midstream
    - Stream bend
  - Waterway
    - Ditch
    - Rapid
  - Spring
    - Hot spring
    - Geyser
    - Sulphur spring
  - Waterfall
    - Cataract
    - Cascade
- Stagnant body of water
  - Lake
    - Lagoon
    - Chain of lagoons
    - Salt lake
      - Intermittent salt lake
    - Chain of intermittent salt lakes
    - Chain of salt lakes
    - Underground lake
    - Intermittent lake
    - Chain of intermittent lakes
    - Glacial lake
    - Crater lake
    - Chain of crater lakes
    - Oxbow lake
      - Intermittent oxbow lake
  - Chain of lakes
  - Pond
    - Salt pond
      - Intermittent salt pond
    - Chain of salt ponds
    - Fishpond
    - Chain of fishponds
    - Horsepond
    - Mere
    - Millpond
  - Pool
    - Intermittent pool
      - Billabong
    - Mud puddle
    - Wallow

Advocacy for External Quality in GIS

Christelle Pierkot[1], Esteban Zimányi[2], Yuan Lin[3], and Thérèse Libourel[1,3]

[1] UMR ESPACE-DEV (IRD-UM2), Montpellier, France
christelle.pierkot@ird.fr
[2] Université Libre de Bruxelles, Belgium
ezimanyi@ulb.ac.be
[3] LIRMM, Montpellier, France
name@lirmm.fr

Abstract. Nowadays, geographical resources (both data and applications) are increasingly being accessible via search engines or web services. As a consequence, users must choose among a set of available resources the ones that best fit their needs. However, users neophytes are currently unable to determine a priori (i.e., before acquisition and use), whether a resource is adequate for its intended usage. Although metadata, if available, allow users to obtain information about *internal* data quality, this metadata is specified in terms of the data producer, who does not know all the intended uses for the resource. This information is not sufficient for users to evaluate the quality of resources in relation to their needs, i.e., the *external* quality. In this paper, we propose a method that takes into account the user profile, the application domain, the requirements, and intended use to assess, a priori, the quality of the resources.

1 Introduction

Nowadays, the uses of geographic information diversify and multiply. One reason for this is that geographical resources (both data and applications) abound and are available, mostly through the Web. However, this accessibility has compounded a significant problem, the assessment of the quality of the resources and their adequacy for the intended usage. In particular, usages within the public domain (e.g., land use planning, environmental monitoring, risk mapping, etc.) require additional vigilance in this respect.

Therefore, users are faced with the necessity to evaluate the *external* quality of the resources, i.e., their adequacy to the particular usage they are intended for. However, this is a problematic situation since this evaluation is based on an objective component, i.e., the *internal* quality declared by the producer, which is not always available. Furthermore, the evaluation is also strongly correlated to the context of use, which includes objective aspects (e.g., hardware, software) but also cognitive aspects associated with users' knowledge and the expression of their requirements.

The findings reported in this paper result from a survey realized among a set of users of geographical information [3]. The survey shows that the majority of users do not know the quality of a spatial resource before using it, mostly because of

C. Claramunt, S. Levashkin, and M. Bertolotto (Eds.): GeoS 2011, LNCS 6631, pp. 151–165, 2011.

the ignorance of the corresponding metadata, or because an evaluation procedure is not available. This results in general user dissatisfaction. Our proposal is therefore to provide users with a "quality assurance" approach.

This paper is structured as follows. Sect. 2, devoted to the state of the art, gives the definitions and principles around the concepts of internal and external quality, and reports about standardization work and related research around the evaluation of external quality. Sect. 3 is devoted to the heart of the proposal. It first presents the metamodel for quality, and then details the proposed evaluation process illustrated by a use scenario. Sect. 4 concludes the paper and defines further areas of research.

2 Related Work in Quality

Traditionally, the producer of a data set is the only responsible for defining and assessing its quality [11]. However, several works (e.g., [4,7,20]) has shown the necessity of considering the users' viewpoint to determine whether some data is fit for its use. This clearly implies a change of perspective where users may take their responsibilities to find the appropriate resources.

In the context of geographic information, [4] further specifies the definition of quality depending on the producer or the user point of view as follows:

- *Internal quality* is the set of properties and characteristics of a product or service which confers on it the ability to satisfy the specifications of its content. It is measured by the difference between the resource which should have been produced and the resource which has actually been produced. It is linked to specifications (e.g., to errors that can be generated during data production) and is evaluated in terms of the producer.
- *External quality* is the suitability of the specifications to the user's requirements. It is measured by the difference between the resource wished for by the user and the resource actually produced. It is linked to the users' requirements and thus varies from one user to the next.

Several criteria have been defined for assessing the *internal quality* of a spatial dataset. These include lineage, geometric, semantic and temporal accuracy, completeness and logical consistency [9]. All these criteria have been widely analyzed and are nowadays defined in several standards described next[1].

The ISO 19113 standard establishes the principles for describing the quality of geographic data thanks to two types of information. *Data quality elements* provide quantitative information such as positional accuracy or completeness. *Data quality overview elements* provide general, non quantitative information such as lineage.

The ISO 19138 standard defines *basic data quality measures* (e.g., error indicator, correct item count, etc.) that can be used to specify a set of data quality

[1] Notice however, that these standards are currently being reviewed as part of a new project that aims to unify and harmonize all of them in an unique document: the ISO 19157 standard.

measures for each element described in the ISO 19113 standard (e.g., number of duplicate feature instances for completeness, number of invalid self-intersect errors for topological consistency, etc.)

The ISO 19114 standard provides a framework for evaluating the quality information of geographic data in accordance with the principles defined in ISO 19113. A *quality evaluation process* is defined to determine the quality result between a dataset and the product specification or the user requirements.

ISO 19115 is the *metadata* standard for geographic information. Fig. 1 describes how the elements of the ISO 19113 and ISO 19114 standards are represented in ISO 19115.

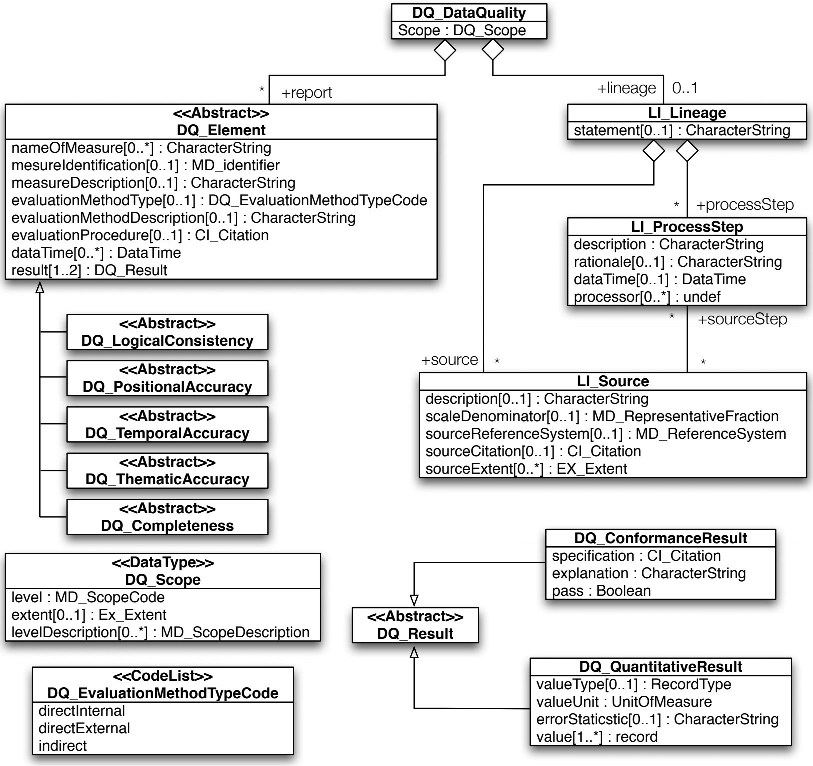

DQ_DataQuality
Scope : DQ_Scope
*
+report
+lineage
0..1
<<Abstract>>
DQ_Element
nameOfMeasure[0..*] : CharacterString
mesureIdentification[0..1] : MD_identifier
measureDescription[0..1] : CharacterString
evaluationMethodType[0..1] : DQ_EvaluationMethodTypeCode
evaluationMethodDescription[0..1] : CharacterString
evaluationProcedure[0..1] : CI_Citation
dataTime[0..*] : DataTime
result[1..2] : DQ_Result
LI_Lineage
statement[0..1] : CharacterString
*
+processStep
LI_ProcessStep
description : CharacterString
rationale[0..1] : CharacterString
dataTime[0..1] : DataTime
processor[0..*] : undef
*
+sourceStep
+source
*
*
LI_Source
description[0..1] : CharacterString
scaleDenominator[0..1] : MD_RepresentativeFraction
sourceReferenceSystem[0..1] : MD_ReferenceSystem
sourceCitation[0..1] : CI_Citation
sourceExtent[0..*] : EX_Extent
<<Abstract>>
DQ_LogicalConsistency
<<Abstract>>
DQ_PositionalAccuracy
<<Abstract>>
DQ_TemporalAccuracy
<<Abstract>>
DQ_ThematicAccuracy
<<Abstract>>
DQ_Completeness
<<DataType>>
DQ_Scope
level : MD_ScopeCode
extent[0..1] : Ex_Extent
levelDescription[0..*] : MD_ScopeDescription
<<CodeList>>
DQ_EvaluationMethodTypeCode
directInternal
directExternal
indirect
DQ_ConformanceResult
specification : CI_Citation
explanation : CharacterString
pass : Boolean
<<Abstract>>
DQ_Result
DQ_QuantitativeResult
valueType[0..1] : RecordType
valueUnit : UnitOfMeasure
errorStaticstic[0..1] : CharacterString
value[1..*] : record

Fig. 1. Quality Information in the ISO 19115 standard

Quality metadata are accessible via the DQ_DataQuality section. Each instance of the class DQ_DataQuality is characterized by a scope that specifies the nature of the target data, in particular the application level and the geographical area . The class DQ_DataQuality is an aggregation of two classes that provide genealogy information (LI_Lineage), and quantitative information such as the precision of the data (DQ_Element). The results of quality measures are available by DQ_QuantitativeResult and DQ_ConformanceResult elements.

However, quality information available in the ISO 191xx series, is typically used to describe the quality of resources from the producer's viewpoint and does not take into account the user's viewpoint.

External quality, and particularly data relevance, is a concept that can be linked to the concept of fitness for use. In the last few years, much research has been done for taking into account external quality [2,7,12,17,20]. [7] points out that properly defining data quality requires information about data usage but also about user requirements. Recently, [10] defines quality as the proximity between data characteristics and needs of a user for a given application at a given time.

Two broad approaches have been proposed in the litterature for determining external quality. One of them is based on the assessment of the risk inherent to the use of inadequate data [2,12]. The other is based on the use of metadata to analyze the similarity between the data produced and the users' needs [7,17,20]. The proposal for assessing external quality that we present in this paper relies on the metadata approach.

3 Assessing External Quality

Fig. 2 is the starting point of our approach for external quality assessment.

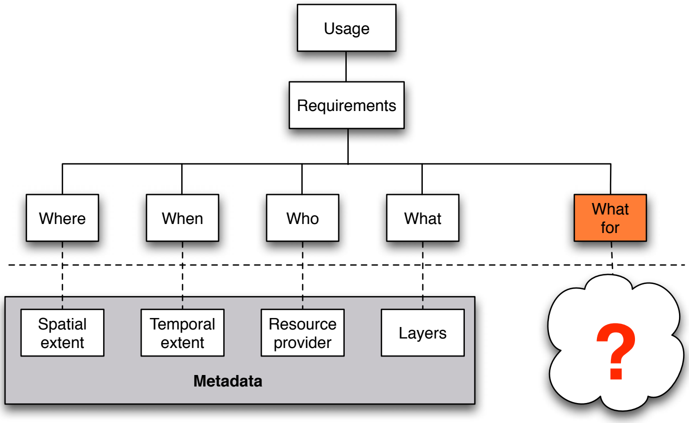


Fig. 2. Specifying usage for selecting geographical resources

When selecting geographical resources, users typically start from a spatial search engine, which relies on metadata to select a set of resources that address the following questions: 1) *Where*, for defining the spatial extent, 2) *When*, for defining the temporal extent, 3) *Who*, for defining the resource provider, and 4) *What*, for defining the layers the user is interested in. What is currently missing is an additional dimension: 5) *What for*, for defining the usage that it is expected for the resources.

Providers of geographical resources are begining to take this last dimension into account. For example, the French Maping Agency IGN allows users to select among all its products by specifying a set of intended usages, e.g., by foot, by bicycle, by car, outdoor activities, touristic information, historical information, etc. However, without the notion of user requirements, the results obtained are too general and it is not possible to evaluate the external quality of the resources.

Therefore, in this paper we propose a metamodel for quality that take into account both the user's and the producer's viewpoint (Sect. 3.1) and describe a process for evaluating external quality (Sect. 3.2).

3.1 A Metamodel for Quality

Fig. 3 gives a general overview of our metamodel for quality. It is composed of two related parts, which allow to define and evaluate the quality of a resource. The left part describes the information about the intended use, such as the domain, the user, and the requirements (user's viewpoint). The right part describes the information about the resource, such as specification and metadata (producer's viewpoint).

The class Resource describes either a geographical Data set or Application. A Ressource is generated by a Producer, that can be either institutional such as National Mapping Agencies, (e.g., IGN in France and Ordnance Survey in

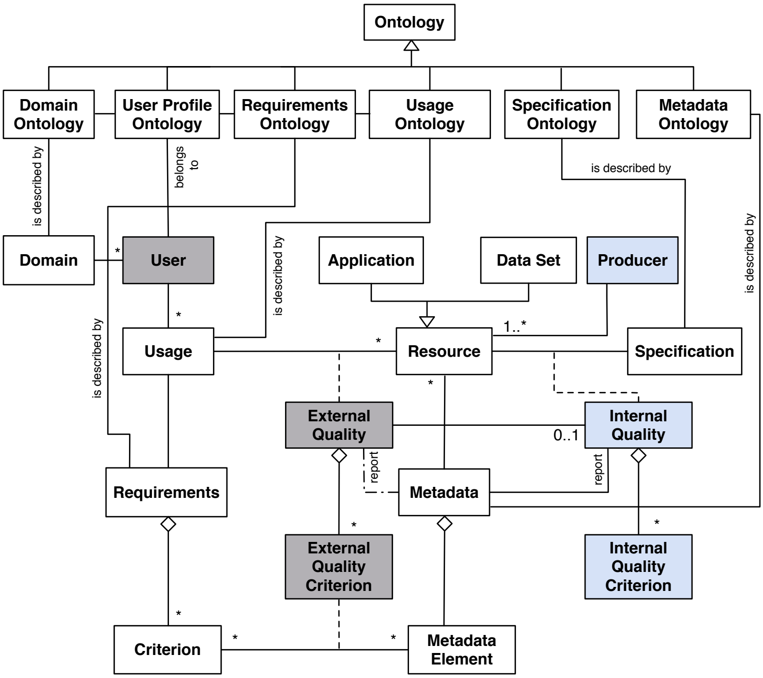


Fig. 3. A metamodel for quality

England) or private such as research team or a user. Geographical data sets are composed of raster data (e.g., France Raster from IGN, OS Landplan Data from Ordnance Survey) or vector data (e.g. BD Topo for France, OS MasterMap Topography for England). An example of a geographical Application is a catalog, which allows users to find the resources they need, either for general a usage (e.g., IGN Géoportail[2], OS OpenData[3], OpenStreet Map[4]), or for a particular application domain (e.g., MAGIC[5] or MDweb[6] for the environmental domain). Other examples of geographical applications are web services that provide users with information that can be added as indicators to map layers that will established beforehand (e.g., Info trafic[7], which shows road trafic, pollution, public works, etc. in Paris). A Resource is linked to one or more Usages and inversely a usage may require one or more resources. A resource may have a Specification, which explain how it was generated (e.g., specifications for BD Topo[8] or for OS MasterMap Topography[9]).

Resources are described by Metadata, which are typically established from a profile. A profile is an aggregation of standardized metadata (e.g., Dublin Core, ISO 19115, Darwin Core) and additional metadata specifically defined for a particular usage, domain, or application.

From the above elements of the metamodel we can assess the Internal Quality of a resource. As defined in Section 2, the Internal Quality measures the adequacy of a resource with its specification. The Internal Quality is an aggregation of several Criterion, which must be defined in accordance with the ISO 19113 principles. The results of the evaluation of the internal quality of a resource is reported as metadata, as recommanded by the ISO 19114 standard.

In the other part of the metamodel, the class Usage describes the general intended use for a particular User in a specific Domain. Examples of usages are biodiversity monitoring, avalanche prediction, or cycling tourism.

A User may require one or more Usages. A User belongs to different profiles, depending on their profession and their expertise on a Domain. For example, in avalanche prediction the same information must be available to the general public, to avalanche experts, and to decision makers [16]. As a User is related to a particular Domain, we can derive associations between an Usage and a Domain. For example, biodiversity monitoring belongs to the environmental domain, avalanche prediction belongs to field of risk management, and cycling tour can be related to tourism. Such domains may be standardized; an example is the

[2] http://www.geoportail.fr/

[3] http://www.ordnancesurvey.co.uk/oswebsite/opendata/

[4] http://www.openstreetmap.org/

[5] http://www.magic.gov.uk/

[6] http://www.mdweb-project.org

[7] http://www.infotrafic.com/home.php

[8] http://professionnels.ign.fr/DISPLAY/000/506/447/5064472/DC_BDTOPO_2.pdf

[9] http://www.ordnancesurvey.co.uk/oswebsite/products/osmastermap/userguides/docs/OSMMTopoLayerUserGuide.pdf

Biodiversity Information Standards[10], which include Darwin Core[11]. Similarly, the ISO 31000 is a family of standards related to risk management.

A Usage may be formalized by a set of Requirements, which are composed by a set of Criterion. Requirements for our previous examples of usages are as follows:

- For biodiversity monitoring, we need phenology information[12], weather information, calendric information, time series of species observations.
- For avalanche prediction, requirements are weather information, snowfall information, altitude gradient,
- For cycling tourism, we need to combine cartographic information with air quality and traffic information services.

As we defined in Sect. 2, External Quality measures the adequacy of a resource with respect to its usage requirements. Currently, assessing external quality depends on quality criterion defined by the producer such as positional accuracy or completeness. This is necessary but not sufficient to accurately evaluate the resources with respect to user requirements. We propose to add to this measure, an independant evaluation by computing some values between the requirements criterion and the metadata elements (see Sect. 3.2 for details).

Further, it is necessary to report the external quality of a resource in the metadata, so that users of the same domain with similar requirements and usages can obtain this information without having to evaluate it. This involves the definition of new metadata fields that do not exist in the standards (e.g., fiability of the producer of the resource) to store this information.

Finally, the main concepts in our metamodel are described by ontologies. There are several reasons for this. First, ontologies in the left part of the schema help the user to better define her objectives, and are used in a search engine such as MDweb [5]. This implies that these ontologies are related so that the links between the different concepts (e.g., domain, profile, usage, requirements, etc.) may be determined. Further, ontology matching is needed for assessing the quality of resources, as described in next section. To achieve these goals, we rely on existing ontologies such as tourism ontologies [18], environment ontology[13], requirements ontologies [13], metadata ontologies[14] [19], and specifications ontologies [1].

3.2 Process for Evaluating the External Quality

In this section we present the process of external quality assessment.We illustrate this by using the scenario of a user who wants to generate a cycling touristic maps for Paris. Notice that the user acts as a prosumer (i.e., a producer-consumer) of

[10] http://www.tdwg.org/

[11] http://rs.tdwg.org/dwc/

[12] Phenology is the study of how periodic plant and animal life cycle events are influenced by seasonal variations in climate.

[13] http://www.environmentontology.org/

[14] Translations (e.g. in OWL) of standard metadata typically used for semantic interoperability.

geographical information, since it aggregates information from multiple sources in order to produce the resource.

Finding available resources corresponds to answering the questions *where, when, who, what,* and *what for* depicted in Fig. 2. It is assumed that users are able to select the spatial extent (e.g. Paris) via a search engine like MDweb [5] which returns a set of resources [15] answering the *where* question. However, it is more difficult to specify the requirements for the *when* (e.g., today, maximum 1 hour, etc.), *who* (e.g., IGN, AirParif, etc.), *what* (e.g., roads, points of interest, etc.) and *what for* (e.g., cycling tour) questions. Thus, the set of resources found by the search engine must be evaluated and refined to give a better result which satisfy all or most of the user needs. This is done with the help of a three-step process as follows:

1. Formalize user requirements and specify the main objectives,
2. Find correspondances between user's requirements and metadata of available resources,
3. Assess the external quality and select the resources that best satisfy user requirements.

We detail next each of these steps.

Step 1. In this step, we must help the user to formalize requirements and to valuate them in order to establish the objectives.

First, the user chooses an application domain among those proposed by the system (e.g. tourism, environment, etc.). This is done thanks to a domain ontology. From this, the system proposes different user profiles within this domain. These user profiles are defined from two elements: the profession and the expertise level. For example, one profession in the touristic domain is that of tour organizer, and the expertise levels in this profession may range from novice to professional.

Following this, the system determines a set of typical usages, based on the domain ontology and the user profile, as well as by interacting with the user when she wants to supply additional information. When the user specifies new usages, the system automatically update the corresponding ontologies with the new information. For example, in the case of setting up a touristic map, the usages belong to the following categories:

- Proposed by the system:
  - Transportation means, i.e., walking, cycling, public transportation, or car tour.
  - Type of interest, e.g., cultural, natural, gastronomic, or sport. Each of them can be further specified, e.g., cultural can be specialized into museums, momuments, historical, etc.
  - Specific constraints, e.g., handicapped needs, children, family, etc.

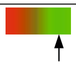
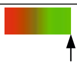
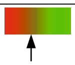
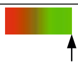
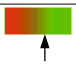
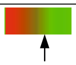
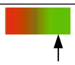
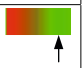
[15] The result is composed of several type of resources whose metadata corresponds with the spatial extent requested.

– New ones specified by the user, e.g., avoid polluted places and congested traffic roads, overall cost, and fiability.

From the combination of these predefined usages, the system determines a set of formalized requirements with the help of the requirements ontology. For the usages specified by the user, the ontology does not contain predefined criteria to formalize them; therefore the system asks the user to propose new criteria for enriching the ontologies. For exemple, the usages "avoid polluted places and congested traffic roads" brings the user to define new criteria such as pollution index et traffic index. For our example of cycling touristic map for Paris, the requirements criteria are thus: positional accuracy (Acc), road network (Roads), orography (Orog), points of interest (PoI), pollution index (PolIx), traffic index (TrafIx), fiability (Fiab), and overall cost (Cost).

These criteria will be then displayed to the user so she can valuate them. In our example, the user wants data with a positional accuracy of at least 10 meters, pollution index with freshness of at most one day in the ATMO scale (defined by French regulations), traffic index with freshness of at most one hour, fiability of 80%, and all of that with a maximal cost of 20 €. Finally, the user must determine the weight of each criteria, which is a value in $[0, 1]$. In our example this would result in Table 1, which shows the requirements criteria and the corresponding user objectives, the latter represented by the desired value and its weight. In the figure the weights are represented graphically, where red corresponds to 0 and green to 1.

Table 1. User objectives

	Acc	Roads	Orog	PoI	PolIx	TrafIx	Fiab	Cost
Objective	≤10 m	Y	Y	Y	≤1 day	≤1 h	≥ 80%	≤ 20 €
Weight								

Step 2. In this step, we must find the correspondences between the user requirements specified in the previous step and the metadata of each individual resource found by the search engine. Since the requirements criteria and the metadata are expressed using formal ontologies, finding the correspondences amounts to an ontology matching problem [8].

To enable interoperability, we suppose that the available metadata of the resources comply with those defined by the ISO 19115 standard. If for a particular application domain, the metadata contained in the standard are not enough, then a community profile[16] should be established to add the missing metadata, in conformance with ISO recommandations.

The heterogeneities between two ontologies may be of several types [8]: syntaxical, terminological, conceptual, and semiotical. We cope here with terminological heterogeneities (i.e., those concerning entity names, such as synonyms

[16] An ISO profile corresponds to an extension and/or a restriction of the ISO 19115 standard by a particular user community.

or when using several natural languages) and conceptual heterogenities (i.e., those concerning differences in modelling of the same domain)[17]. An example of the latter concerns the pollution index. The French ATMO pollution index is composed by the following pollutants: sulphur dioxide, dust particles, nitrogen dioxide, and ozone, while the European one (CiteAIR) includes many more such as carbon monoxide and hydrocarbons.

We cope with terminological and conceptual heterogeneities using *string-based* techniques, which allow to find the entity names that correspond to each other (exactly or similarly). This is the case for criteria already existing in the ontologies and whose terms have been defined according to those defined in the ISO 19115 standard. For example, the criterion positional accuracy corresponds to the class DQ_PositionalAccuracy in ISO 19115. Other techniques such as those based on *linguistic resources* (synonyms, hyponyms, etc), the *taxonomy-based techniques* (using subsumption links) or those based on *upper level and domain-specific ontologies* (commonsense knowledge or domain knowledge) are used for finding the correspondances between the criteria added by the user and the metadata.

The techniques above must be sometimes associated to a global strategy. This is the case, e.g. when finding a correspondance between the cost criterion introduced by the user and a metadata element of the ISO 19115 standard. We use then two matching techniques, i.e. a linguistical one to find the synonyms (price, fee, charge, expense, etc.) and a terminological one to compute the similarity between the names. Matcher composition methods (sequential or parallel composition) allows several matching algorithms to be combined. For our exemple, the final result of the composition of both algorithms gives the element MD_Ditributor.MD_StandardOrderProcess.fees as corresponding to the cost criterion.

Notice that there may be a many-to-many correspondance between the criteria and the metadata, since several criteria could correspond to a metadata element and conversely, several metadata elements may be aggregated into a single criterion. For example, the road criterion relative to the thematic layer can be found in several metadata elements such as MD_Identification.abstract or MD_Keywords.keyword. Notice that it may be the case that there is no correspondence between the criteria and the metadata. This is taken care progressively by the systems through the update of ontologies and metadata profiles taking into account users' input.

Following this, we must determine the correspondences at the instance level, i.e. between the criteria values and the metadata values. However, the values may be defined in different units. A typical example concerns costs which can have different type (e.g. expressed in euros or in dollars). As another example, the ATMO index has a value domain from 0 to 10, while that of CiteAir has values from 0 to 100. We cope with this problem using matching methodes of

[17] Syntaxical heterogeneities are supposed to to be solved previously to the evaluation process. Semiotical heterogeneities are difficult to cope in an automatic way, since they depend on the user interpretation of an application domain.

type *language based* (use of a dictionnary) or *constraint based* (use of the internal structure of the entities, i.e. type, value domain, cardinality, etc.).

Finally, we obtain a set of correspondances resulting from several matching strategies. From this set, we build a correspondance matrix between the available resources and the user requirements. For our example, this results in Table 2, which establishes how the resources Res1–Res7 satisfy the requirements. For example, Res1 has a positional accuracy of 1 m, provides information about roads, about traffic index with freshness of less than one day, has fiability of 80%, and is free, while information about points of interest can be found in resources Res6 and Res7. Notice the values in the columns of the matrix must be translated into the same units (e.g., hours for traffic information); this is necessary for determining the utility functions in the next step.

Table 2. Correspondance matrix $\mathcal{M}_R$ relating resources with requirements

	Acc (m)	Roads	Orog	PoI	PolIx (hours)	TrafIx (hours)	Fiab (%)	Cost (€)
Res1	1	Y				24	0.80	0
Res2	1	Y				1	0.90	10
Res3	1	Y			1	0.25	0.95	100
Res4	5				24		0.75	0
Res5	1	Y	Y				0.60	100
Res6	10	Y	Y	Y			0.70	0
Res7	5		Y	Y			0.90	0

Step 3. Starting from the correspondance matrix established in the previous step, a multicriteria decision-aid method must be applied to choose a set of resources that optimizes the user objectives. In our case, given the resources $\mathcal{R} = \{R_1, \ldots, R_m\}$, the alternatives $\mathcal{A} = 2^{\mathcal{R}}$ are the subsets of $\mathcal{R}$ and the criteria $\mathcal{C} = \{C_1, \ldots, C_n\}$ correspond to the user requirements.

As stated by [15], several multicriteria decision analysis methods are available and selecting the one to be used depends on the decision problem at hand. Further, such methods must be customized to our particular setting. Among the numerous methods that have been proposed, we will focus on the family of Multi-Attribute Utility Theory (MAUT) methods [14].

First, we need to compute a correspondance matrix $\mathcal{M}_A$ for the alternatives. For an alternative A composed of resources $\{R_1, \ldots, R_k\}$, this amounts to aggregate the values of the correspondance matrix $\mathcal{M}_R$ for the resources R_i. How this is done depends on the kind of criterion to be considered. In our example, for accuracy we take the maximum value (the least accurate), since when combining resources of varying accuracy, the accuracy of the result is given by the least accurate resource. The reason for this is that, e.g., a data set with 1 m accuracy can be converted to an accuracy of 5 m, but it is not always possible to do the reverse conversion. Similarly, for binary functions (e.g., roads) the maximum is also be chosen. However, for cost the sum must be used, since the cost of an alternative A is the sum of the cost of its components resources. Finally, for fiability the average can be used. Thus, for an alternative A composed of resources $\{R_1, \ldots, R_k\}$, its correspondance to criterion C_i is given by

$$C_i(A) = \Theta_i(a_{ji}), \text{ for } j = 1, \ldots, k$$

where Θ_i is an aggregation function (e.g., min, max, sum, average) defined by the user and a_{ji} is the cell of the correspondance matrix $\mathcal{M}_R$ relating resource R_j and criterion C_i. Table 3 shows the correspondance matrix for three alternatives.

Table 3. Correspondance matrix $\mathcal{M}_A$ for alternatives (only three of them are shown)

	Acc (m)	Roads	Orog	PoI	PolIx (hours)	TrafIx (hours)	Fiab (%)	Cost (€)
{Res2, Res4, Res7}	5	Y	Y	Y	24	1	0.85	10
{Res3, Res7}	5	Y	Y	Y	1	0.25	0.93	100
{Res1, Res4, Res6}	10	Y	Y	Y	24	24	0.75	0
...	...	...	...	...	...	...	...	...

Then, we must define a utility function $g_i : \mathcal{A} \rightarrow Y \in \mathbb{R}$ for each criterion C_i. Such function expresses how well an alternative A satisfies the user objectives for criterion C_i. A utility function has typically a range in $[0, 1]$ and must take into account whether the value of the criterion must be minimized (e.g., cost) or maximized (e.g., fiability). In our case, considering Table 3, the utility functions for binary criteria (e.g., roads) take value 0 or 1, while the utility functions for the other criteria are given by

$$g_i(x) = \begin{cases} \exp(\frac{-x^2\rho_1}{\sigma^2}) & \text{for criteria to be minimized} \\ 1 - \exp(\frac{-x^2\rho_2}{\sigma^2}) & \text{for criteria to be maximized.} \end{cases}$$

In the above formulas, σ is the threshold value of the objective stated by the user (e.g., 20 € for cost), ρ_1, ρ_2 are functions of μ given by $\rho_1 = -\ln(\mu)$ and $\rho_2 = -\ln(1-\mu)$, and μ is the utility value at σ (e.g., 0.8). The parameter μ, which can be customized by the user, determines the distinguishability between a resource that satisfies an objective at the threshold value (e.g., with cost of exactly 20 €) and the resource that best satisfies the objective (e.g., with cost of 0 €). Fig. 4 shows the utility functions when the user wants to minimize (left) or maximize (right) a criterion with $\sigma = 20$ and $\mu = 0.8$.

For example, the left function states that a resource with cost 20 € has a utility value of 0.8 but another resource with cost 0 € is 20% better since it has a value of 1.0. Table 4 shows the values of the utility functions g_i for several alternatives.

Finally, the global multi-attribute utility function must be determined by taking into account the utility functions g_i and the weight w_i of the criteria as expressed by the user in Table 1. First, we normalize the weights w_i by defining $\lambda_i = \frac{w_i}{\Sigma_{j=1}^n w_j}$ in order to ensure that $\Sigma_{i=1}^n \lambda_i = 1$. Then, the utility of an alternative A is given by

$$U(A) = \Sigma_{i=1}^n \lambda_i g_i(A).$$

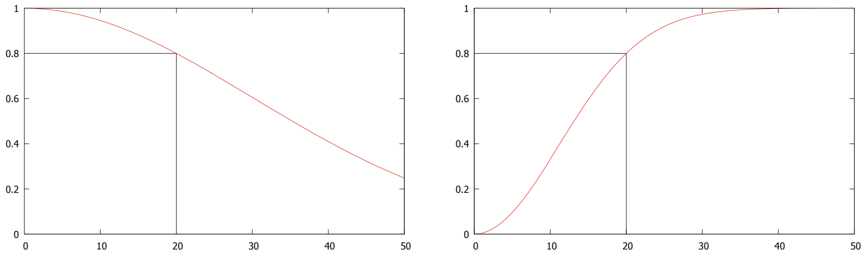


Fig. 4. Utility functions for minimizing (left) and maximizing (right) an attribute with $\sigma = 20$ and $\mu = 0.8$

Table 4. Utility values for the alternatives (results are rounded to two decimal places)

	Acc	Roads	Orog	PoI	PolIx	TrafIx	Fiab	Cost	U(A)
$\{\text{Res2}, \text{Res4}, \text{Res7}\}$	0,95	0,80	0,80	0,80	0,80	0,80	0,84	0,95	0,84
$\{\text{Res3}, \text{Res7}\}$	0,95	0,80	0,80	0,80	1,00	0,99	0,88	0,00	0,76
$\{\text{Res1}, \text{Res4}, \text{Res6}\}$	0,80	0,80	0,80	0,80	0,80	0,00	0,00	1,00	0,74
...	...	...	...	...	...	...	...	...	...

The rightmost column in Table 4 shows the utility value for some alternatives.

By applying the function U above to all alternatives $A \in \mathcal{A}$ we can rank them in decreasing order, so if two alternatives A_1 and A_2 are such that $U(A_1) > U(A_2)$, this means that A_1 satisfies better than A_2 the user objectives.

The result of this step is then a ranked list of alternatives, each alternative being a collection of resources. In our case (cf. Table 4), the best alternative is the one composed by resources Res2, Res4, and Res7, where road and traffic information are taken from Res2, pollution from Res4, and orography and points of interest from Res7, at a total cost of 10 €. This alternative has as utility value of 0.84 and it satisfies all user objectives. A less optimal alternative is the one composed by resources Res3 and Res7, which has a utility value of 0.76. Although this alternative is better than the previous one concerning the objectives for pollution index, traffic index, and fiability, it does not satisfy the objective for cost. Finally, the third alternative composed by resources Res1, Res4, and Res6, although being the best possible for price (0 €), it does not satisfy the objectives for accuracy and fiability and thus, it has an utility value of 0.74. Notice that in the case that no alternative meets all requirements criteria, the ranked list of choices constitutes the best compromise for optimizing the user objectives.

4 Conclusions

We argued in this paper that it is necessary to provide users who access geographical resources with a quality assurance method. We emphasized the fact that both internal and external quality must be taken into account. *Internal quality* concerns the producer's viewpoint and establishes the correspondence of a resource with respect to the specifications. On the other hand, *external quality* concerns the user's viewpoint and establishes the adequacy of a resource with respect to the usage it is intended for.

We adopted a general framework based on a metamodel for quality. In our view, the interest of this metamodel is to emphasize the importance of some knowledge that remains mostly implicit. Making this knowledge explicit is the foundation on which the evaluation process of external quality is built. Starting from a set of formalized user requirements, the evaluation process uses a multi-criteria decision-aid method for establishing a ranked set of resources that best satisfies the requirements.

The research work presented in this paper can be pursued in several directions. First, we need to build ontologies and metadata profiles for other application domains we are interested in, especially environmental and risk management. These ontologies and profiles can be progressively refined in an automatic way by taking into account the criteria added by users. Another issue concerns the automatization of the computation of external quality for resources that are obtained on the fly. Yet another direction consists in storing the result of external quality assessment so that this information can be used by the system for the filtering process, when a user of the same domain, profile, and usage looks for resources.

More generally, we intend to study the role of the evaluation of the quality of resources on the quality of the overall project or organization in which its use is carried out. Further, in our metamodel, a usage is defined with respect to the profiles of users or the project in which they are involved, and therefore, the context is fixed. This reduces the problems related to hardware, software, etc. However, it is necessary to generalize the metamodel to be able to describe usages that span across users and projects. Finally, we wish to achieve the operational implementation of our proposal within the existing platform MDweb [5].

References

1. Abadie, N., Mechouche, A., Mustière, S.: OWL based formalisation of geographic databases specifications. In: Proceedings of the 17th International Conference on Knowledge Engineering and Knowledge Management, Poster, Lisbon, Portugal (October 2010)
2. Agumya, A., Hunter, G.: Fitness for use: Reducing the impact of geographic information uncertainty. In: Proceedings of the URISA Anual Conference, Charlotte, NC, USA (1998)
3. Conte-Tisnerat, Y., Ali, H.E., Gasc, F., Heridi, H.: Qualité externe des données, ontologie des usages. Technical report, UM3, LIRMM, Projet Tutoré, Master TSAD SIIG3T (2010)
4. David, B., Fasquel, P.: Qualité d'une base de données géographique: concepts et terminologie. Technical report, IGN, Bulletin d'information n.67 (1997)
5. Desconnets, J.-C., Libourel, T., Clerc, S., Granouillac, B.: Cataloguing for distribution of environmental resources. In: Proceedings of the 10th AGILE Conference on Geographic Information Science, Aalborg, Denmark (2007)

6. Devillers, R., Jeansoulin, R. (eds.): Fundamentals of spatial data quality. Geographical Information Systems series, ISTE (2006)
7. Devillers, R., Jeansoulin, R.: Spatial data quality: Concepts. In: [6], ch. 2, pp. 31–42
8. Euzenat, J., Shvaiko, P.: Ontology matching. Springer, Heidelberg (2007)
9. Guptill, S., Morisson, J.L. (eds.): Elements of Spatial Data Quality. Pergamon Press Inc., Oxford (1995)
10. Gutiérrez, C., Servigne, S.: Métadonnées et qualité pour les systémes de surveillance en temps-réel. Revue Internationale de Géomatique 19(2), 151–168 (2009)
11. Harding, J.: Vector data quality: A data provider's perpective. In: Devillers, R., Jeansoulin, R. (eds.) [6], ch. 8, pp. 141–160
12. Hunter, G., Bruin, S.D.: A case study in the use of risk management to assess decision quality. In: Devillers, R., Jeansoulin, R. (eds.) [6], ch. 14, pp. 271–282
13. Jureta, I.J., Mylopoulos, J., Faulkner, S.: A core ontology for requirements. Applied Ontology 4(3-4), 169–244 (2009)
14. Keeney, R.L., Raiffa, H.: Decisions with Multiple Objectives: Preferences and Value Trade-Offs. Cambridge University Press, Cambridge (1993)
15. Laaribi, A., Chevallier, J., Martel, J.: Spatial decision aid: A multicriterion evaluation approach. Comput., Environ. and Urban Systems 20(6), 351–366 (1996)
16. Parent, C., Spaccapietra, S., Zimányi, E.: The MurMur project: Modeling and querying multi-represented spatio-temporal databases. Information Systems 31(8), 733–769 (2006)
17. Pierkot, C.: Vers un usage éclairé de la donnée géographique. In: Actes de l'atelier Qualité des Données et des Connaissances de EGC 2010, Hammamet, Tunisie (2010)
18. Prantner, K., Ding, Y., Luger, M., Yan, Z., Herzog, C.: Tourism ontology and semantic management system: State-of-the-arts analysis. In: Proceedings of the IADIS International Conference WWW/Internet 2007, Vila Real, Portugal, pp. 111–115 (October 2007)
19. Schuurman, N., Leszczynski, A.: Ontology-based metadata. Transactions in GIS 10(5), 709–726 (2006)
20. Vasseur, B., Jeansoulin, R., Devillers, R., Frank, A.: External quality evaluation of geographical applications: An ontological approach. In: Devillers, R., Jeansoulin, R. (eds.) [6], ch. 13, pp. 255–270

Multi-criteria Geographic Information Retrieval Model Based on Geospatial Semantic Integration

Walter Renteria-Agualimpia[1,2] and Sergei Levashkin[1]

[1] Laboratory of Intelligent Processing of Geospatial Information,
Centre for Computing Research (CIC), National Polytechnic Institute (IPN),
Av. Juan de Dios Bátiz s/n, 07738, Mexico D.F., Mexico
[2] Advanced Information Systems Lab (IAAA),
Computer Science and Systems Engineering Department,
Aragon Institute for Engineering Research (I3A),
University of Zaragoza,
Edificio Ada Byron, María de Luna, 1, E-50018 Zaragoza, Spain
walter.renteria.agualimpia@gmail.com

Abstract. The geospatial semantic web development requires search mechanisms to overcome the syntactic comparisons and perform semantic analysis in order to retrieve information conceptually similar to the searched one by the user. This would allow reducing the risk of return empty results *no match found* when there is no exact correspondence between the query and the information available in data repositories. In this work, we describe an information retrieval model to integrate a semantic criterion with geospatial criteria in a no homogenous vector space. The criteria represent the dimensions of this space; these dimensions are weighted in function of the user's preferences or his/her profile. The integration is based on a mathematical expression to evaluate the relevance of each result. We present a system that implements the geospatial semantic approach proposed to retrieve information from the domain of cultural tourism, specifically museums. The results show the advantages of integrating geospatial and semantic criteria taking into account user profiles to offer more customized (personalized) service.

Keywords: geographic information retrieval, spatial semantics, semantic similarity, multi-criteria integration, user profile, service personalization.

1 Introduction

The systemic approaches to the geographic information retrieval are aimed at returning results on geographical objects that satisfy a query. However, many of known systems lack of a semantic component that allows returning results somewhat relevant to the queries, when the retrieval system is expected to produce such results that are most similar to the searched one. A difficulty in achieving this goal can be solved by means of geospatial semantic processing integration. Although many systems today offer different approaches to the search, there is no standard way of integration to a unified outcome taking into account all relevant criteria. Another factor hindering the integration is the nature of the data which are not structured, i.e. heterogeneous.

C. Claramunt, S. Levashkin, and M. Bertolotto (Eds.): GeoS 2011, LNCS 6631, pp. 166–181, 2011.

In this work we propose a search system that returns results rather than, or in addition to, exact values in response to user queries and user preferences (user profile). The system utilizes an ontology-based scheme knowledge organization that provides a semantic component to geospatial processing, a relational database as a repository of geographic data for spatial processing, and a novel system integration of results with a weighting configurable by the user in a no homogenous multidimensional space. The information retrieval model uses the criteria of semantic similarity proposed in [7] and spatial analysis. Spatial analysis consists of combining Euclidean distance and proximity analysis of museums surrounding the geographic object searched for by the user.

2 Related Work

Information retrieval is a complex area, this is further complicated when required to overcome the syntactic matching. There have been efforts to retrieve semantically the results, in some of them (e.g. [3], [4]) metadata, natural language processing and semantic strategies are used to expand queries. However, these efforts are not very efficient, because the restrictions on the number of query terms and lack of standards regarding the domain terminology, especially the concepts that form the conceptual structures such as ontologies.

The work [11] uses other approaches like rule-based and data-driven methods to search for better results. This article presents several heuristics to access data resources and direct the recovery thereof.

Undoubtedly, one of the strongest approaches is based on the use of ontologies [8], [9], [16]. In [5], they propose a system where each annotation is assigned a weight that reflects the relevance of the analysis in order to determine the meaning of a document. The system computes weights based on the frequency of occurrence of the instances in each document.

Another strategy also proposes a use of weights either configured by the user or weights calculated by other methods, one of them is based on concept frequency analysis. An example of this strategy is TF–IDF [17]. It is a simple method, often used in information retrieval and text mining. It employs weight as a statistical measure; these weights are used to compute the importance of a word in a document belonging to a corpus.

The systems that use formal spaces are described by Paul Churchland and Peter Gardenfors, who propose a theory of space similarity of concepts by stating that "concepts are regions of similarity spaces that are somehow realized in the brain" [1]. Based on this theory, Raubal in [12] proposes a formal representation of cognitive semantics to describe a methodology and formalize conceptual spaces as sets of quality dimensions with a geometrical structure. Such spaces can be used for knowledge representation based on the mathematical theory of vector spaces and a statistical standardization method.

In addition to the knowledge representation, many aspects of the real world can be captured in such systems as GIS: Spaces full of discrete spatial objects or even continuous measurement of several different properties or themes within a concrete spatial region [9]. Besides these geographical properties, there are properties with a higher level of abstraction that allow us to relate one geographic object to another semantically in order to get places (i.e. museums) that are very similar to others according either to definitions of recognized international organizations (UN, UNESCO, ICOM, etc.) or to opinion of people who use them.

Nowadays, although there are strategies for integrating criteria or preferences [20], they however do not take into account the semantic aspect to combine it with geospatial analysis. The work [24] proposes a system for mobile devices to assess the geographical relevance, which includes many aspects such as: Spatial (where), temporal (when), topics of objects (what or which) and motivational user. Part of this proposal focuses on the analysis of spatial proximity, using buffers.

The work [23] proposes an ontology-based model to recover and weigh the geographic information from unstructured repositories, using geographic and topological relations extracted from heterogeneous information sources. And then, weigh the topological, geographic, and semantic results by means of a method denominated *iRank. iRank* evaluates the relevance of each document (DG), given a query QG by formula (1):

$$RelInt = \frac{RelCon(Cq,Cd) + RelGeo(Gq,Gd) + RelTopoly(Tq,Td)}{geographic\ sources\ number} \quad (1)$$

In [25], they propose spatiotemporal ontological relevance model (STORM), which uses a three-dimensional grid to represent the degree of importance of geospatial documents according to spatial, temporal, and semantic similarity criteria. This approach is presented as an interactive interface for three-dimensional visualization that can complement the relevant evaluation systems based on lists.

Christopher Jones *et al.* [2], propose a way to retrieve information based on integration of measures. They developed a hybrid distance, combining a hierarchical distance measure with Euclidean distance, which are weighted by pre-set weights. The expression (2) to evaluate the relevance ultimately is the sum of the conceptual distance (hierarchical), and geometric distance:

$$TSD(q,c) = w_e ED(q,c) + w_h HD(q,c) \quad (2)$$

Where W_e is the weight for the Euclidean distance and W_h is the weight for the hierarchical distance.

Our work recaptures some of the exposed ideas. The main goal of this research is to develop a model for geospatial semantic integration in the domain of cultural tourism in order to retrieve information about Cultural Points of Interest (CPI), especially museums. Developed herein computing system can recover types of museums (e.g. painting museum) even if their names are not stated syntactically or explicitly as museums of painting.

3 Geospatial Semantic Integration

3.1 Semantic Analysis

One of the main contributions of this work is the use of semantic analysis as a complement to enhance the spatial information retrieval. In [18], they show that using semantic and ontological structures approach enriches the geographical and topological relations. The semantic richness is achieved by using ontologies, which provide a formal framework to represent the shared knowledge of a domain [13].

The importance of using semantic analysis lies in the possibility of extending the results to others that are similar in order to reduce (or even avoid) the empty results (*no match found*) and increase the number of results that could be of the user's interest due to their semantic similarity. This is especially true in the area of geographic information retrieval, where there may be geographical objects that satisfy a query or even items that are relevant if we admit a small margin of error, that is, if we have not exact correspondence with the query, there are similar objects to retrieve. These objects are similar or conceptually satisfying the query under a small margin of error, and the results can serve to reduce gaps.

In this paper we have use hierarchies [15] as a particular case of ontologies. The hierarchies are considered as conceptual structures and used to represent domain knowledge studied (our case study is *museums of the downtown Mexico City*). The *confusion theory* [7] allows us to add semantic analysis component to traditional geospatial analysis. Figure 1 below illustrates the main hierarchical structure used. It has been designed into the Protégé environment as an OWL file [19]. In Protégé, we are defined classes and relationships, in this case, the relationship is "is-a" to relate the classes and subclasses, e.g. *painting* museum is an *art* museum.

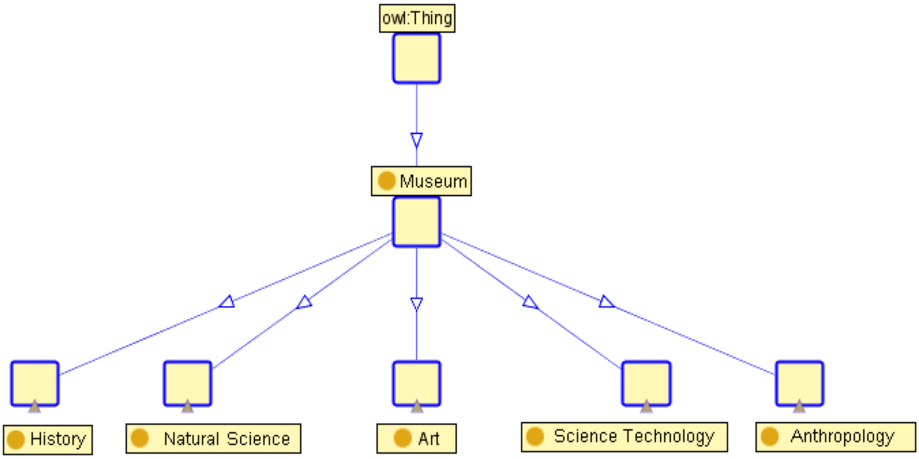


Fig. 1. Hierarchy of museums according to information provided by ICOM [14]

3.1.1 Similarity Measure

There are several proposals and models to measure the similarity between conceptual entities. Some of them are based on the features of the entities [21]. Others compute similarity through other strategies such as those proposed by Rodriguez and Egenhofer [22]. A few of these strategies are similar to that used in this work, because they use conceptual structures to determine the semantic similarity between two entities on the conceptual structure.

We base on the fact that there may be geographic objects or entities satisfying a query with a small margin of error that can be controlled in order to return results rather than or in addition to exact values in response to user queries. We use *confusion theory* [7] to measure the similarity between geographic objects. In the following, we present a short description of this theory:

Definition 1. A **hierarchy** H of an element set M is a tree whose root is M and if a node has children then these form a *partition* of their parent.

Definition 2. For an element set M, a **hierarchy** H of M is a tree of nodes; each node is either an element of M or a set of symbolic values v_i, for i = 1, … , m, where $v_i \propto E_i$, and $\{E_1,E_2,\dots, E_m\}$ is a partition of M.

Example: For M = {Painting, Art, Sculpture, Museum, Arqueology, "Museo Mural Diego Rivera", "Museo del Templo Mayor", Anthropology, Museo del Palacio de Bellas Artes, Museo Nacional de San Carlos,}, the hierarchy is:

H_1 = **{**Museum $\propto$ {Art Museum $\propto$
{Painting Museum $\propto$
{"Museo Mural Diego Rivera",
"Museo del Palacio de Bellas Artes"}
Sculpture Museum $\propto$ {"Museo Nacional de San Carlos"}
}
Anthropology Museum $\propto$ { Arqueology Museum $\propto$
{"Museo del Templo Mayor"}
}
}

The hierarchy groups M into smaller sets of the same symbolic values. The former definition v $\propto$ M is used, where the symbolic value v represents the set M (see Figure 2).

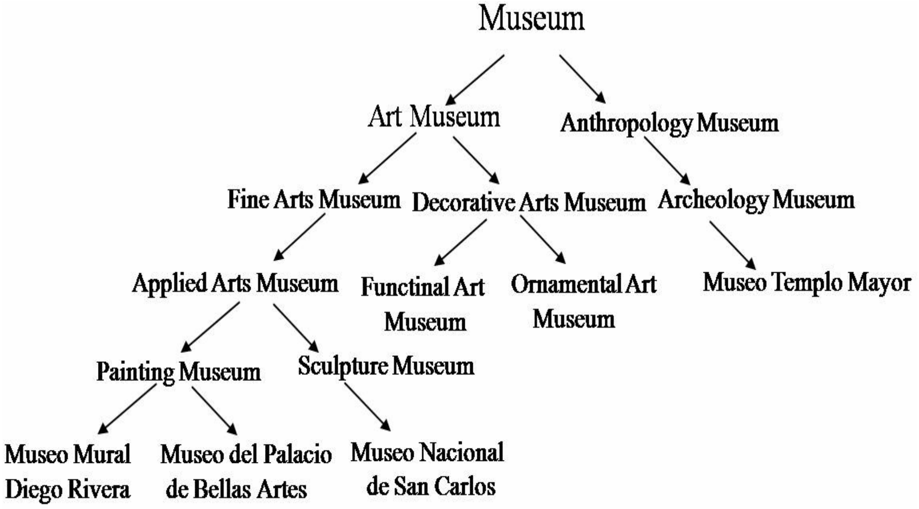


Fig. 2. Museums Hierarchy**:** A set of elements in the hierarchy (nodes and symbolic values)

For example, one asks for an art museum and the answer is “Mural Diego Rivera Museum”. Is it a mismatch? Yes, but very small, since “Mural Diego Rivera Museum” is a painting museum and painting is a kind of art. An example of a still smaller error: One asks for “Templo Mayor Museum” and an Anthropology museum is recommended to her. Can we classify or measure such mismatch? We can do it using hierarchies (see Figure 2). In this section, we measure the mismatch (called *confusion*) when one symbolic value is used instead of another (the intended or correct value).

If $r, s \in H$, then the ***confusion*** in using r instead of s, written *conf(r, s)*, is:

- *conf (r, r)* = *conf (r, any ascendant_of (r))* = 0;
- *conf (r, s) = 1 + conf (r, father_of(s)).*

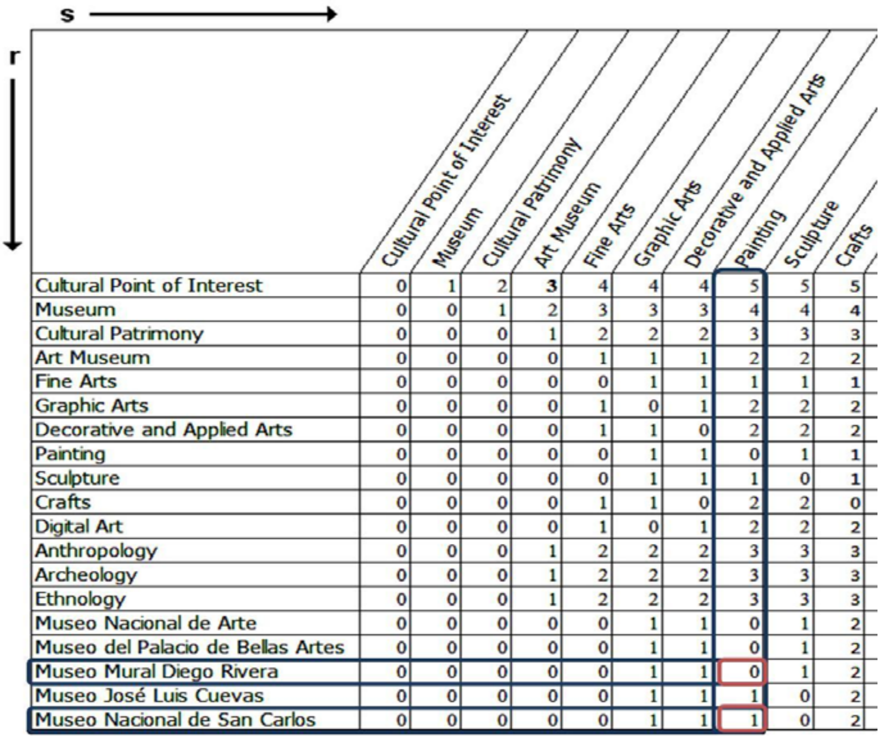
To measure *conf*, count the *descending* links from r (the replacing value) to s (the replaced or intended value). Finally, without loss of generality, it is possible to define *conf (r,s)* = [1 + *conf (r, father_of(s))*] / *height* (*H*). In this case, 1 ≥ *conf (r,s)* ≥ 0.

The main idea is to find similarity concept by measuring the semantic distance between local concepts of the ontology.

In this work, it is assumed the confusion value as the measurement of semantic similarity, i.e, a smaller value of *consufion* means major semantic similarity.

In the table below, we show an example of *confusion* in using r instead of row s column for Museums Hierarchy.

Table 1. Confusion in using row *r* instead of column *s* for Museums Hierarchy (Fig. 2)



s →

r ↓

	Cultural Point of Interest	Museum	Cultural Patrimony	Art Museum	Fine Arts	Graphic Arts	Decorative and Applied Arts	Painting	Sculpture	Crafts
Cultural Point of Interest	0	1	2	3	4	4	4	5	5	5
Museum	0	0	1	2	3	3	3	4	4	4
Cultural Patrimony	0	0	0	1	2	2	2	3	3	3
Art Museum	0	0	0	0	1	1	1	2	2	2
Fine Arts	0	0	0	0	0	1	1	1	1	1
Graphic Arts	0	0	0	0	1	0	1	2	2	2
Decorative and Applied Arts	0	0	0	0	1	1	0	2	2	2
Painting	0	0	0	0	0	1	1	0	1	1
Sculpture	0	0	0	0	0	1	1	1	0	1
Crafts	0	0	0	0	1	1	0	2	2	0
Digital Art	0	0	0	0	1	0	1	2	2	2
Anthropology	0	0	0	1	2	2	2	3	3	3
Archeology	0	0	0	1	2	2	2	3	3	3
Ethnology	0	0	0	1	2	2	2	3	3	3
Museo Nacional de Arte	0	0	0	0	0	1	1	0	1	2
Museo del Palacio de Bellas Artes	0	0	0	0	0	1	1	0	1	2
Museo Mural Diego Rivera	0	0	0	0	0	1	1	0	1	2
Museo José Luis Cuevas	0	0	0	0	0	1	1	1	0	2
Museo Nacional de San Carlos	0	0	0	0	0	1	1	1	0	2

With the results of Table 1 we can process the query: "Painting Museum". The results are shown in Table 2 according to hierarchy of museums (Fig. 2).

Table 2. Example of a query processing by *confusion*

value similarity ***conf*(r, Painting Museum)** **i.e. the value of the error ε**	**result set = r**	**Instance of**	
0	Museo del Palacio de Bellas Artes	Painting	Cluster 0
0	Museo Nacional de Arte	Painting	
0	Museo Mural Diego Rivera	Painting	
1	Museo José Luís Cuevas	Sculpture	Cluster 1
1	Museo Nacional de San Carlos	Sculpture	
2	Museo Franz Mayer	Decorative Art	Cluster 2
2	Laboratorio de Arte Alameda	Graphic Art	
3	Museo Nacional de las Culturas	Ethnology	Cluster 3
3	Museo de la Ciudad de México	Ethnology	

The table shows that it is desirable to use additional criteria to rearrange or refine the results in each cluster. For example, the results with the same value of confusion (i.e. semantically equal results) are located at different distances from the user.

3.2 Spatial Analysis

This part of the work is composed of two well-known operations, the calculation of distance between two objects and analyzing geographic proximity that is carried out by space operations, including the buffering operation performed by the calculation of spatial density as section 3.2.2 describes.

3.2.1 Geometric Distance

It is really not simple to measure the distance between two places of a city, because there are many factors to consider to move from a point A to point B: In big cities like Mexico City, it would require more travelling time to go from A to B and not in the opposite direction. Such factors as the pedestrian bridges allow shortening the way to people but not cars. Driving a car you should consider other cars, the availability of parking in addition to the journey time to walk afterwards. These and other factors make the distance from one point to another no longer symmetrical and trivial. For the sake of simplicity, however, we choose a simple metric – the Euclidean distance. Thus, we use the linear distance between the user position and every museum *m* that meets the query.

3.2.2 Spatial Density

The traditional analysis of proximity does not take into account the different perceptions that people have for the concept of proximity and closeness: For one person, a place which is at a distance less than a mile is near, while for people with mobility difficulties this represents an enormous distance.

In order to show an example of proposed model, the spatial density of museums are pre-computed using a simple algorithm, the density is stored in a field of spatial database – the number of museums that are located within an area A_m and within the radius $r_{density}$ and with the center at geographic coordinates (x_m, y_m) of museum *m*. This set of values represents a dimension of the integration space of the criteria (see in the following section 3.3). In future work, we can let the user choose the search area or radius. For now, the radius is set at one kilometer. The primary target is allowing the users decide if interests to them to go to a zone with much culural wealth (i.e. high concentration of museums).

3.3 Space for Criteria Integration

We propose the integration of different criteria for information retrieval defining a space called *Space for Criteria Integration*. The main objective is to combine a no homogenous component (i.e. the semantic analysis) with geospatial analysis, so that they complement each other giving back more satisfactory results. We believe this is a way to approach the human beings answer and make recommendations on several alternatives, taking into account several criteria.

Space for Criteria Integration has the following properties[1]:

(***i***) Each dimension is used to mapping a recovery criterion.
(***ii***) Each dimension or criterion is independent on the other.
(***iii***) Each dimension is a set so that the values closer to zero are most appropriate to satisfy a query.

The space graphic representation appears in Figure 3.

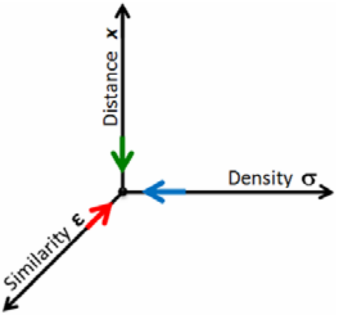


Fig. 3. Values close to zero (space's origin) are more satisfactory to the user's query

In this way, a museum *m* that satisfies the user's query Q, is composed of a tuple of values $m(\varepsilon, \sigma, x)$.

[1] Further properties of this space will be described in a subsequent paper.

- The value of similarity ε is defined as: If $\varepsilon \to 0$, then the museum *m* is very similar to the searched one and if $\varepsilon = 0$ the museum meets the exact query. Note that the similarity dimension ε is not homogenous, i.e. the similarity values are not constants for all domains: For instance, a similarity value $\varepsilon = \varepsilon_1$ in tourism domain is diferent of the same value $\varepsilon = \varepsilon_1$ in biological domain.
- The distance value x is defined as the distance between the geographic coordinates of the user making the query and the geographic coordinates of the museum (x_m, y_m): If $x \to 0$, then the museum *m* is very close to the user.
- The value of density σ' is defined as the number of museums that are located within an area A_m of a museum *m*. To satisfy (***iii***), we apply the following transformation $\sigma = 1 - \dfrac{\sigma'}{\max\left(Density_m\right)}$: If $\sigma \to 0$, then the *m* museum has high density of cultural points of interest, i.e. *m* museum has many other museums around it. The function *max ( )* returns the maximum value of the set of values in $Density_m$.

(***iv***) Each dimension has a significant degree, determined by a weight *w,* which provides the users an option to decide what criteria are most important to them in the query. The set of weights *w* is a user profile. This profile is defined by the configuration that the users provide to their preferences. Some of the main ideas to catch the preferences arise from the area of multicriteria analysis decision [10]. It is important to note that the user does not have to be an expert to set these weights: Users only need to "equalize their preferences", i.e. attenuate or give priority to each criterion according to their wishes. The "*equalizer preferences*" is a model component that captures the user's needs or desires, so he/she should only answer the simple question:

How important is each criterion for you?

That is, the users have to indicate the degree of importance that they attach to each criterion according to their preferences to define a *profile* as shown in Figure 4.

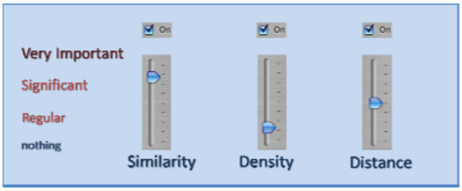


Fig. 4. The degree of importance of each criterion is the weight for each dimension

(*v*) When a query is about a museum, their $m(\varepsilon, \sigma, x)$ values are combined to form a single point in *Space for Criteria Integration*: Each outcome of the query will be represented by a point in the space.
(***vi***) The relevance value *R* of each result ("each point") is calculated as shown in the expression (3):

$$R(c_i, w_i) = 1 - \frac{\sum_{i=1}^{n} c_i^{\,w_i}}{n} \tag{3}$$

Where *n* is the number of criteria c_i and w_i are the weights set by the user through the equalizer, i.e. the user profile.

Finally, *R* determines the results that come closest to the geographic object into query, so that: If $R \to 1$, then the result is more satisfactory or more relevant for a query; Figure 5 shows graphic representation of *R*.

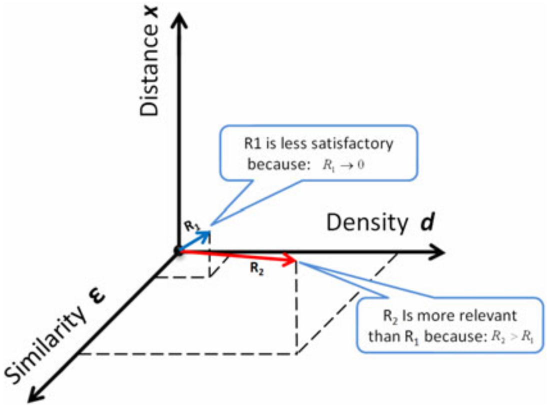

Distance x
R1 is less satisfactory because: $R_1 \to 0$
Density d
R1
R2
Similarity ε
R2 Is more relevant than R1 because: $R_2 > R_1$

Fig. 5. The *R* value determines the degree of relevance of each result, so that R_1 is less relevant because it is closer to origin, but R_2 is a more satisfactory result to be greater than R_1

Finally, the most relevant geographical objects to answer a query are those with a tuple of values closer to the origin simultaneously.

3.4 System Architecture

The proposed model of information retrieval and integration is shown in Figure 6. The main phases of this model are the conceptualization of the domain of Cultural Points of Interest (CPI) "especially museums", this conceptualization takes as data sources provided by ICOM [14], and governmental organizations such as CONACULTA-Mexico and INEGI-Mexico. We also use the dictionaries of INEGI-Mexico (National Institute of Statistics, Geography and Informatics).

Aiming to give flexibility to the model, we choose to separate the data into two parts: "two Databases DB". One part "DB Conceptual CPI" serves to represent the semantics through an ontology of museums without including in this the instances and second part “DB Geographic CPI” consists of the description of each instance (every museum in the historical center of Mexico City), so that each of the DB Museum is geographically referenced by nodes in the ontology. If one wishes to add other instance, then he/she won’t affect the ontology nodes “semantic part”.

DB Conceptual or hierarchy "is-a" (as special case of ontologies) is the source for semantic analysis by calculating the semantic similarity of concepts – confusion (section 3.1.1), which is then integrated with geospatial analysis (calculation of distance and proximity analysis), using the S*pace for Criteria Integration* to respond queries about geographic objects (museums in the historic center of Mexico City).

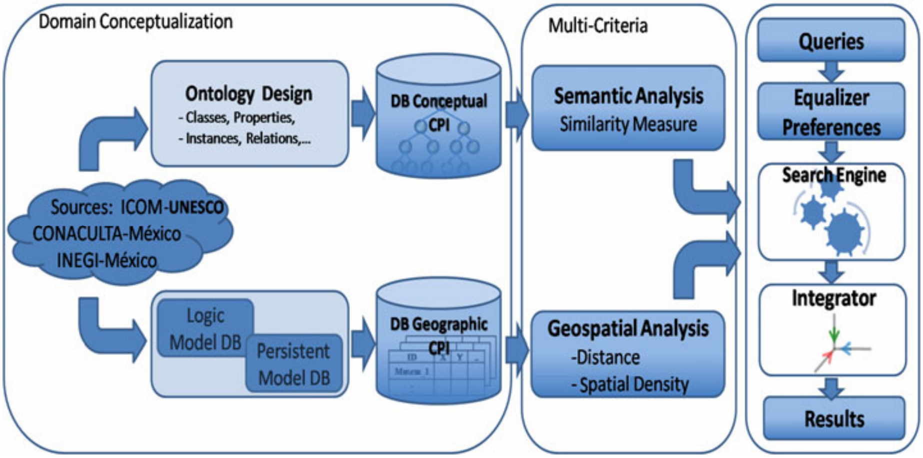


Fig. 6. Information retrieval system architecture, comprising the conceptualization of the domain into two databases of Cultural Points of Interest – CPI (Conceptual and Geographic DB of CPI) for the semantic and geospatial analysis and integrate the results in the *Space for Criteria Integration*

4 Results

The results are presented by comparing the answers obtained without using the integration with the multi-criteria combination and user preferences. For instance, if the search is only performed based on the similarity criterion, i.e. using the confusion measure, it can sort the results that have different values. A disadvantage is that one cannot establish an order on elements of the same class, i.e. the nodes of hierarchy that are children of one parent, unless it is ordered to work on hierarchies. To rearrange instances of a class requires an additional criterion (see Table 2). That is why we use geospatial criteria as well.

On the other hand, if only spatial criteria are used growing disadvantage and loss of semantic richness that the similarity analysis through ontologies provides. The results, presented in the following, are not only integrated the spatial and semantic criteria, but also allowed the participation of the user, who assigns a degree of importance to each criterion.

For example, Table 3 shows the query results using the criteria individually for query = “Painting Museum” close to “subway station “Hidalgo”. The results are presented by similarity (second column), depending on the distance (third column), according to the density (number of Cultural Points of Interest (CPI) in the fourth column). However, the main drawback is to decide on the best CPI, how to rank them from high to low relevance: For example, "*Museo Historia y Cultura Naval*" is very close to the location specified by the user, but it is not semantically similar or "*Museo Mural Diego Rivera*" is exactly a painting museum, but it is farther than 500 meters. These examples explain how difficult is to make a decision when there are no unified criteria.

Table 3. Query results. Query = "Painting Museum" close to "subway station "Hidalgo".

Museum	ε	d	x	Description
Museo Nacional de Arte	0.14	0.06	0.249	Very similar, very dense and near
Museo del Palacio de Bellas Artes	0.14	0.06	0.119	Very similar, dense and very near
Museo Mural Diego Rivera	0.14	0.14	0.539	Very similar, very dense and intermediate
Museo José Luis Cuevas	0.29	0.06	+1.00	Similar, dense and far
Museo Nacional de San Carlos	0.29	0.14	+1.00	Similar, very dense but far
Museo de la SHCP, Antiguo Palacio del Arzobispado	0.29	0.09	0.416	Similar, dense and intermediate
Museo Nacional de Arquitectura	0.29	0.06	0.119	Similar, but very dense and very near
Laboratorio de Arte Alameda	0.43	0.14	0.489	Similar, very dense and intermediate
Museo del Ejercito	0.43	0.06	0.324	Not similar, dense and near
Museo Historia y Cultura Naval	0.43	0.06	0.217	Not similar, dense and very near
Museo Franz Mayer	0.43	0.09	0.231	Not similar, dense and very near
Museo Interactivo de Economía	0.43	0.06	0.373	Not similar, dense and near
Museo del Templo Mayor	0.43	0	+1.00	Not similar, dense and far

On the other hand, Table 4 shows the results for a user who decides to give more importance to the semantic aspect, preferring museums with high similarity to the query, while choosing small weight for the spatial density and relatively high for the distance, i.e. Wε=0.9, Wd=0.2 and W_X=0.7.

Table 4. Results of R "relevance values" using criteria and weights

Museum	$R(\varepsilon,W_{\varepsilon}, d,W_{d}, x,W_{x})$	Description
Museo del Palacio de Bellas Artes	0.9	**Most Relevant** Very similar to query, dense and close to location selected by the user
Museo Nacional de Arquitectura	0.89	
Museo Nacional de Arte	0.86	
Museo Historia y Cultura Naval	0.85	
Museo Franz Mayer	0.82	
Museo de la SHCP, Antiguo Palacio del Arzobispado	0.8	
Museo del Ejercito	0.8	
Museo Interactivo de Economía	0.8	
Mural Diego Rivera	0.71	
Laboratorio de Arte Alameda	0.69	
Museo del Templo Mayor	0.66	
Museo José Luis Cuevas	0.63	
Museo Nacional de San Carlos	0.56	

Clearly the most satisfactory results are the Cultural Points of Interest: "*Museo del Palacio de Bellas Artes", "Museo Nacional de Arquitectura"*, and *"Museo Nacional de Arte"*. Indeed, they are: 1) Semantically most similar; 2) *Fine Arts Museum* and simultaneously 3) Closest to the location specified by the user "between 100 and 250 meters", having high other museums density near them.

Another experiment also shows how users weigh the criteria through their preferences. In this example (Table 5), the user wants to visit Cultural Points of Interest somewhat similar to CPI queried, but mostly CPI that are very close to the location specified in the query, discarding the density, i.e. W_{ε}=0.5, W_d=0.1 and W_x=0.9.

The user query is: "Archeology Museum" close to "subway station "Zócalo".

Table 5. Query results. Query = "Archeology Museum" close to "subway station "Zócalo".

Museum	ε	D	x	Description
Museo Nacional de las Culturas	**0.14**	0.15	0.185	Very similar, dense and very near
Museo Ex-Teresa Arte Actual	**0.43**	0	0.189	Not similar, very dense and very near
Museo de la Autonomía Universitaria	**0.14**	0.12	0.226	Very similar, dense and near
Museo del Templo Mayor	**0**	0.05	0.269	Satisfies exactly, very dense and near
Museo José Luis Cuevas	**0.43**	0.1	0.317	Not similar, dense and near
Museo de la Ciudad de México	**0.14**	0.07	0.420	Very similar, dense and intermediate
Museo de la Luz	**0.43**	0.12	0.469	Not similar, dense and intermediate
Museo Nacional de Arte	**0.43**	0.1	+1.00	Not similar, dense and far
Museo Franz Mayer	**0.43**	0.12	+1.00	Not similar, dense and far

In this table, we can observe that if the system only takes into account a criterion to rank results (e.g. distance criterion), then the results can be less satisfactory, because of, although "*Museo Nacional de las Culturas*" is the closest one (185 meters), it is not a museum semantically similar to the query, i.e. it is not a museum of *"Archeology"*. Note that in fourth place "*Museo del Templo Mayor*" appears, being an *Archeology Museum* and locating at some 269 meters from user's position.

Now if we compare the results obtained by integrating criteria clearly the most satisfactory is the Cultural Point of Interest: "*Museo del Templo Mayor"*, because it is not only similar, but exactly *Archeology Museum*, and simultaneously, among all the museums, it is the nearest *Anthropology Museum* "between 200 and 300 meters" from the subway station "Zócalo". Additionally, first museums like *Archeology Museum* (i.e. *Ethnology Museum*) are shown, because the *Archeology* and *Ethnology Museums* are sub-class-of the *Anthropology Museums*, and then the rest of results are displayed according to the relevance of museums. In conclusion, first museums more similar and closer simultaneously appear (Table 6).

Table 6. Results of R "relevance values" using criteria and weights

Museum	$R(\varepsilon, w_\varepsilon, d, w_d, x, w_x)$	Description
Museo del Templo Mayor	0.93	**Most Relevant** exactly satisfies the query
Museo de la Ciudad de México	0.82	
Museo de la Autonomía Universitaria	0.80	
Museo Nacional de las Culturas	0.78	
Museo Ex-Teresa Arte Actual	0.77	
Museo José Luis Cuevas	0.67	
Museo de la Luz	0.63	
Museo Nacional de Arte	0.60	
Museo Franz Mayer	0.58	

It is important to note here that the user's preferences (a set of the weights w_i) provide more accurate results (instances or geographical objects), approaching better the user's expectations.

The rank of results obtained earlier by the user has given preference to geographical objects very similar and very close to the location indicated in the user's query but a bit less important to those which have the highest density of cultural tourist places.

The experiments and obtained results discussed in this section make us believe that our methodology approaches the way in which the human beings recommend options in function of several criteria.

5 Conclusions

This work describes an approach to geographic information retrieval based on semantic geospatial integration. It combines spatial and semantic analysis to produce results that are more relevant and more satisfactory according to user preferences than the results based on individual criterion.

It is important to note that the interaction of the user through *equalizer preferences* makes the results more accurate, since each criterion is weighted according to the user's preferences and keeps the queries more customized to obtain results based on the user's profiles.

A space for criteria integration is proposed to allow representing and analyzing the integration of criteria, while searching geographic information by individual preferences. The silence results *no match found* are reduced and the accuracy of the results is controlled using the semantic similarity in conjunction with the geospatial processing.

We face a none-trivial problem of integration semantic and geospatial processing: Indeed, some data are processed in qualitative context and others in quantitative. We present an approach to solve this problem by using hierarchies ("is-a") as a special case of ontologies. In addition, the *confusion theory* is used to enable the integration of the concepts with spatial analysis in a hierarchical data structure.

The obtained results demonstrate that it is possible to integrate criteria into a single criterion using the geospatial semantic approach. We believe that this is close to the way of how people make decisions considering many criteria.

Acknowledgments

The authors are grateful to the Centre for Computing Research (CIC), National Polytechnic Institute (IPN), and the Mexican National Council for Science and Technology (CONACYT) for their support.

References

1. Gärdenfors, P.: Conceptual Spaces: the Geometry of Thought. MIT Press, Cambridge (2000)
2. Jones, C.B., Alani, H., Tudhope, D.: Geographical Information Retrieval with Ontologies of Place. In: Montello, D.R. (ed.) COSIT 2001. LNCS, vol. 2205, pp. 322–335. Springer, Heidelberg (2001)
3. Petras, V., Gey, F.C., Larson, R.R.: Domain-Specific CLIR of English, German and Russian Using Fusion and Subject Metadata for Query Expansion. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 226–237. Springer, Heidelberg (2006)
4. Delboni, T.M., Borges, K.A., Laender, A.H., Davis, C.A.: Semantic Expansion of Geographic Web Queries Based on Natural Language Positioning Expressions. Transactions in GIS 11(3), 377–397 (2007)
5. Castells, P., Fernández, M., Vallet, D.: An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval. IEEE Transactions on Knowledge and Data Engineering 19(2), 261–272 (2007); Special Issue on Knowledge and Data Engineering in the Semantic Web Era
6. Salton, G., McGill, M.: Introduction to Modern Information Retrieval. McGraw-Hill, New York (1983)
7. Levachkine, S., Guzman-Arenas, A.: Hierarchy as a New Data Type for Qualitative Variables. Expert Systems with Applications: An International Journal 32(3), 899–910 (2007)
8. Guarino, N.: Formal ontology and information systems. In: Proceeding of FOIS 1998, Trento, Italy, pp. 3–15. IOS Press, Amsterdam (1998)
9. Egenhofer, M., Burns, H.: Visual Map Algebra: a directmanipulation user interface for GIS. In: Proc. Working Conference on Visual Database Systems (1995)
10. Wang, H.F.: Multicriteria Decision Analysis, From Certainty to Uncertainty, pp. 166–180 (2003)
11. Egenhofer, M.: Interaction with Geographic Information Systems via Spatial Queries. Journal of Visual Languages and Computing 1(4), 389–413 (1990)
12. Raubal, M.: Formalizing Conceptual Spaces. In: Varzi, A., Vieu, L. (eds.) FOIS 2004, pp. 153–164. IOS Press, Amsterdam (2004)
13. Gruber, T.: What-is-an-ontology?, `http://www-ksl.stanford.edu/kst/what-is-an-ontology.html`
14. ICOM, International Council of Museums (ICOM) (2001), `http://icom.museum/hist_def_eng.html` (visited February 2009)

15. Levachkine, S., Guzmán-Arenas, A.: Hierarchies measuring qualitative variables. In: Gelbukh, A. (ed.) CICLing 2004. LNCS, vol. 2945, pp. 262–274. Springer, Heidelberg (2004)
16. Buscaldi, D., Rosso, P., García, P.: Inferring geographical ontologies from multiple resources for geographical information retrieval. In: Proceedings of 3rd Int. SIGIR Workshop on Geographic Information Retrieval, SIGIR, Seattle, pp. 52–55. ACM Press, New York (2006)
17. Salton, G., McGill, M.: Introduction to Modern Information Retrieval. McGraw-Hill, New York (1983)
18. Mata, F.: Geographic Information Retrieval by Topological, Geographical, and Conceptual Matching. In: Procceding of Second International Conference on Geospatial Semantics, GeoS 2007, Mexcio City, Mexico (2007)
19. The Protégé Ontology Editor and Knowledge Acquisition System (2008), `http://portege.stanford.edu`
20. Bernard, L., Kanellopoulos, I., Annoni, A., Smits, P.: The European geoportal - one step towards the establishment of a European spatial data infrastructure. Computers, Environment and Urban Systems 29, 15–31 (2005)
21. Tversky, A.: Features of similarity. Psychological Review 84(4), 327–352 (1977)
22. Rodríguez, A., Egenhofer, M.: Comparing geospatial entity classes: an asymmetric and context-dependent similarity measure. International Journal of Geographical Information Science 18(3), 229–256 (2004)
23. Mata, F.: Recuperación y Ponderación de Información Geográfica desde Repositorios No Estructurados Conducidas por Ontologías, México D.F., Presentada en el Centro de Investigación en Computación CIC-IPN para obtención del grado de Doctor en ciencias de la Computación (2009)
24. Reichenbacher, T.: Geographic Relevance in Mobile Services. In: Proceedings of the Second International Workshop on Location and the Web (LocWeb 2009), Boston, Massachusetts (April 2009)
25. Hobona, G., James, P., Fairbairn, D.: An evaluation of a multidimensional visual interface for geographic information retrieval. In: The CIKM Workshop on Geographic Information Retrieval, pp. 5–8. ACM Press, New York (2005)

A Description Logic Approach to Discover Suspicious Itineraries from Maritime Container Trajectories

Paola Villa and Elena Camossi

European Commission Joint Research Centre,
Institute for the Protection and Security of Citizen,
Ispra, Varese, Italy
{paola.villa,elena.camossi}@jrc.ec.europa.eu

Abstract. About 90% of the world's cargo is transported in maritime containers, but less than 2% is physically inspected by custom authorities. The standard method to handle this problem consists in document-based risk analysis and route-based risk indicators to target anomalies. In this paper, we exploit a logic based approach to identify suspicious patterns in container itineraries. Specifically, we present an ontology to explicitly formalize the knowledge of the maritime container domain and a formalisation of two suspicious movement patterns to enable their discovery in a knowledge base. The formalisation can be extended to support the discovery of other itinerary patterns. Furthermore, the approach we present can be the basis for future development towards the formalisation and search of patterns in itineraries.

1 Introduction

The analysis of moving object trajectories has become a common practice in Risk Analysis. For example, to support maritime security and fight commercial frauds, Maritime Surveillance authorities employ risk indicators for the evaluation of the trajectories of cargos, ships, and vessels, targeting high-risk consignments of goods and proceeding with costly physical inspections only when needed. Established risk analysis is mainly document-based, i.e., it typically involves a number of risk indicators based on customs declarations given by the importers, such as the type of goods transported, their declared origin etc.

A noticeable improvement in risk analysis has been achieved by also considering the itineraries followed by maritime containers used for goods trading, which are known to be an important risk factor.

The route-based risk indicators developed here take into account spatial information such as the ports where a container has been loaded and discharged, the transshipment ports, the actual itinerary followed, etc., and have been successfully tested.

In the area of risk analysis, the ConTraffic system developed at the JRC is an innovative service platform devoted to the monitoring of commercial containers and

C. Claramunt, S. Levashkin, and M. Bertolotto (Eds.): GeoS 2011, LNCS 6631, pp. 182–199, 2011.

risk assessment within the Maritime Surveillance domain. ConTraffic integrates document-based analysis that relies on customs declarations with the development of route-based risk indicators. The project is carried out in the framework of mutual assistance between EU customs, in collaboration with the European Anti-Fraud Office (OLAF).

As illustrated in Fig. 1, the system continuously gathers container movement data from a number of on-line sources. Then, these data are processed to remove ambiguities and discrepancies, and finally they are loaded into the ConTraffic Data Warehouse to be analysed off line. The dataset covers a significant percentage of the worldwide shipping traffic activities done by the main carrier companies (the estimated coverage is about the 30%): currently about 1 billion movement records are recorded, giving information on about 12 million containers travelling between 35,000 locations. The system fosters the identification of suspicious container cargo consignments that deviate from standard and expected behaviors, and has been successfully tested for cases of false declaration of origin. However, route-based risk indicators have a greater computational complexity, because of the massive amounts of information analysed and because geographical information is inherently complex to process.

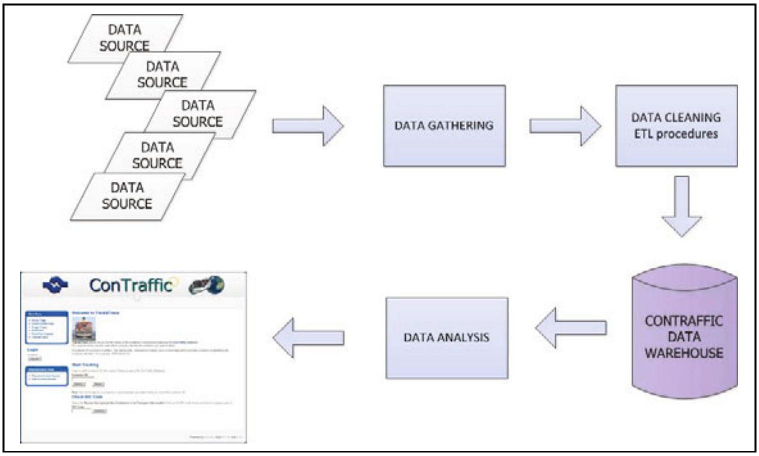


Fig. 1. Overview of the ConTraffic system

In this paper we exploit logic based approaches to develop *semantic route-based risk indicators (SemRI)* for the identification of suspicious cargo consignments. SemRIs target anomalies through the analysis of the itineraries followed by the containers, exploiting not only the spatial aspect of movements but taking into consideration also the explicit knowledge of the application domain. Specifically, SemRIs rely on the asserted and inferred knowledge of *events* occurring to a container, i.e., handling activities performed during the shipping (e.g., discharged, transshipped), and to vessels, which characterise the vessel itinerary. This work is carried out in the application domain of interest to ConTraffic, whose main objective is the discovery of commercial fraud and duties circumvention in the context of Europe; therefore, remarkable itineraries are those involving import-export shipping and cross border transportation, which are mainly delivered by sea.

Recent advancements in reasoning technologies for spatial data demonstrate the efficacy of the exploitation of semantics to enhance the analysis of spatially enabled information such as Moving Objects itineraries, that are not taken into account by the risk analysis currently applied by Maritime Surveillance authorities. Specifically, in this paper we present the Maritime Container Ontology (MCO), to explicitly formalize the knowledge of the maritime container domain, including a semantic-driven representation of container itineraries, integrating the explicit semantics of the events. Moreover, we develop a set of logical axioms to implement semantic route-based risk indicators for the discovery of anomalous and inconsistent *itinerary patterns*.

An example of suspicious patterns discoverable with this approach is the *Loop* [11], where the container is transshipped on another vessel that goes back to the originating port before reaching the destination. Another suspicious pattern, which we refer to as *Unnecessary_Trans*, is used to misdeclare the origin of a container to profit of advantageous duties [11]. This involves an unnecessary transshipment to a vessel that goes to the same destination, therefore the container appears to be coming from the source port of a different vessel (i.e., the vessel that performs just the second part of the trip). Indeed, this type of anomalous itinerary may be successfully discovered by integrating the knowledge of the locations where the events occur and the events' semantics. Semantic driven risk indicators may complement the analysis done by Maritime Surveillance authorities, and the more sophisticated approaches that adopt route-based risk indicators, paving the way to the development of enhanced risk indicators in the container security domain.

The remainder of this paper is organized as follows. In Section 2 we present previous research exploiting the semantics of moving objects trajectories. In Section 3 we give an overall description of the MCO, which formalizes the application domain knowledge of maritime containers. In Section 4 we present the description logic formalisation of suspicious container itineraries that are the core of SemRI. In Section 5 we discuss the potential development of the approach we are proposing and its shortcomings, outlining future research direction. Conclusions and open issues are reported in Section 6.

2 Related Work

In the area of Geographical Information Science, the logic-based semantic representation of trajectories has been recently exploited to reason on touristic itineraries. A preliminary work is given in [4], where the authors have proposed a data pre-processing model to add semantic information to trajectories in order to facilitate data analysis.

However, the most popular way to exploit semantics hidden in the trajectories so far is by developing ad-hoc models and software, considering a mathematical representation of geographic information in order to extract more meaningful patterns, but without relying on a formal system to infer new knowledge: for example, in [2], the authors propose a framework for semantic trajectory knowledge discovery. In particular, they integrate trajectory sample points to the geographic information which is relevant to the application. In [7], the authors define a semantic trajectory data mining query language with several functions to select, pre-process, and transform trajectory sample points into semantic trajectories at higher abstraction levels, in order to allow the user to extract meaningful, understandable, and useful patterns from trajectories. Also in [13], the authors introduce an extensible trajectory annotation model, which is oriented towards the notion of episodes and allows a clear separation of semantic and physical trajectory information. In [1], the authors describe a reverse engineering framework for mining and modeling semantic trajectory patterns.

The latter approaches are close to the one of [6], where the authors present a general framework for conceptually modeling trajectory patterns. The proposed model is an extension of the conceptual model proposed by Spaccapietra and others in [19] for modeling trajectories of moving objects from a semantic point of view. The authors extend this model to support semantic trajectory patterns, that are extracted from aggregated stops and moves of trajectories.

In [17] and [20], we have two analyses that differ from previous researches: in the first one the authors propose an exploratory statistical approach to detect patterns of movement suspension without the necessity of spatial or temporal information about the moving entities and their spatial context. In [20] the authors show a software that integrates semantic and spatial reasoning in SWI-Prolog.

Finally, another way to exploit semantics to define formally trajectories in space and time is the algebraic one: in [21], an algebraic data type is used for a semantic-based representation of spatio-temporal trajectories that integrates the thematic, spatial, temporal and spatio-temporal dimensions at the data representation and manipulation levels.

Our approach differs from [1,2,6,7] and [13] mainly because they describe geographic information using mathematical concepts (i.e., stops and moves), but without exploiting them within a true formal background: the involved algorithms and query languages are ad-hoc. In this respect, the work in [4] is closer to ours, because it relies on description logic framework in order to extract significant patterns.

Moreover, in [17] and [21] the objective to extract patterns exploiting a formalism is similar to ours, but their results are based on different mathematical fields: respectively, statistics and algebra. The approach of [20] is logic based and may open new horizon to our research.

3 The Maritime Container Ontology (MCO)

In this section we give an overall description of the Maritime Container Ontology (MCO) we developed in OWL [8]. In Fig. 2, we report an excerpt of MCO encompassing the main concepts and the roles of interest for the discovery of maritime container patterns. In particular, since we are focusing on containers travelling by sea, the details of other types of transportation are not reported. However, to explain in more detail the framework in which we are developing our work, in Fig. 3 and 5 we report the top-level concepts of the domain (e.g., *Moving object*, *Itinerary*, *Moving Object Event*), showing how MCO extends them through domain specific concepts (e.g., *Container*, *Container Itinerary*, *Container Event*), and which are the most relevant roles between them.

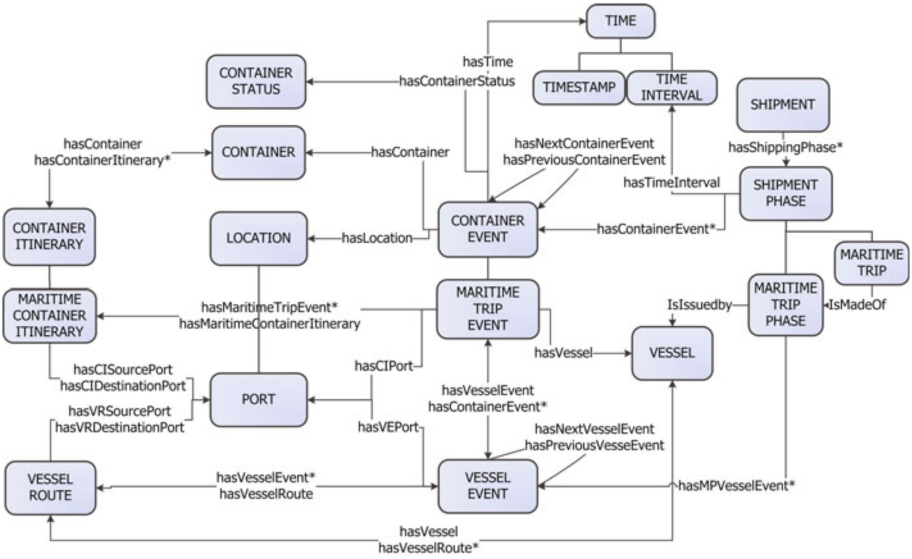


Fig. 2. Maritime Container Ontology (MCO) excerpt

In the ontology diagrams shown in this section, MCO concepts are represented by rectangles with rounded corners; top-level concepts are in bold with darker background. Concept generalizations (i.e., IS-A relationships) are depicted with straight lines (let's assume they go from low-level to top-level concepts), while roles are represented by labeled directed arrows. Starred labels (label*) are one to many relationships. Arrows with double heads represent a role and its inverse. Dashed arrows are roles defined between top-level concepts that need to be further specialized in sub-concepts. For example, *hasEvent* from *Itinerary* to *Moving Object Event* is specialized by *hasContainerEvent* in *Container Itinerary* to link, with appropriate restrictions, *Container Events*; similarly *hasSourceLocation* and *hasDestination* in *Itinerary* are restricted by *hasVRSourcePort* and *hasVRDestinationPort* in *Vessel Event* to link vessel events and *Port*s.

In the following, we first introduce the design for containers and shipments; further, we formalise container events, itineraries, vessels and routes. Finally, we outline the main phases of the process we adopt for the population of MCO with the ConTraffic dataset.

3.1 Containers and Shipments

In the Maritime environment modelled by MCO, every container used for good shipping is identified by a unique international identification code, i.e., the BIC code[1], according to the International Organization for Standardization (ISO) 6346 standard [12]. Analogously, every MCO container is modeled as a unique instance of concept *Container*. Every container belongs to a carrier, i.e., a *Shipping Company* or a leasing company, which leases the container to a carrier.

Shipments delivered via container are formalized by the entity *Shipment* (see Fig. 3), which is further characterized by the *Bill of Lading* of the consignment that includes the references to the containers used for the transportation, the good *Manufacturer*, the *Shipping Company* that handles the shipping, and a *Consignee*. In our formalization we are interested in import-export activities, therefore each shipment is split into three main phases: export and pre-export (*Pre-trip*), the intra-customs trip performed by sea (*Maritime Trip*), import and final consignment (*Post-trip*). Each of them may be further subdivided to describe the activities in which export, import and transport are organized (see Fig. 3).

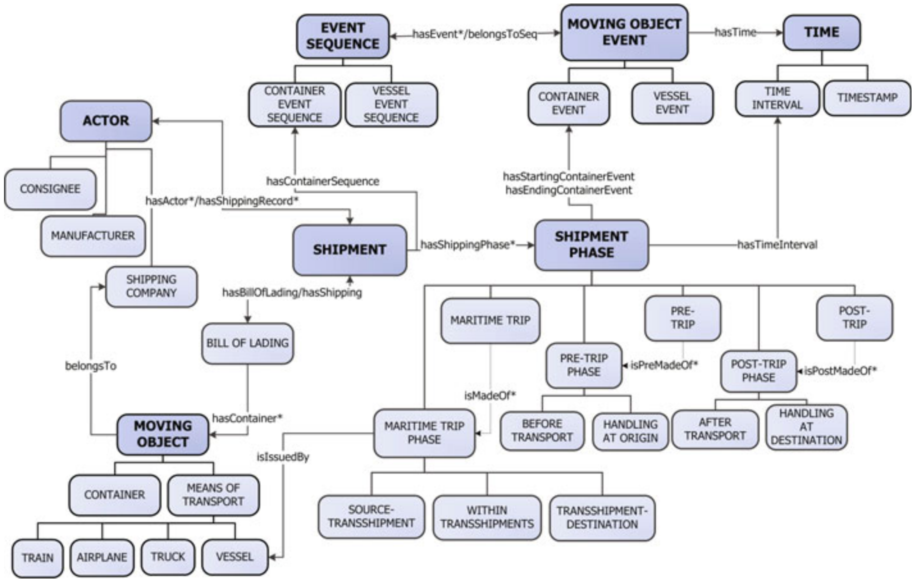


Fig. 3. MCO Shipments

3.2 Container Events

Container Events describe any deed a shipping company undertakes on a container, especially during export and import phases of a shipping (e.g., Loaded to vessel X, Discharged at port Y). Events occurring during the transportation are

[1] BIC codes are assigned by the *Bureau International des Containers et du Transport Intermodal* (BIC).

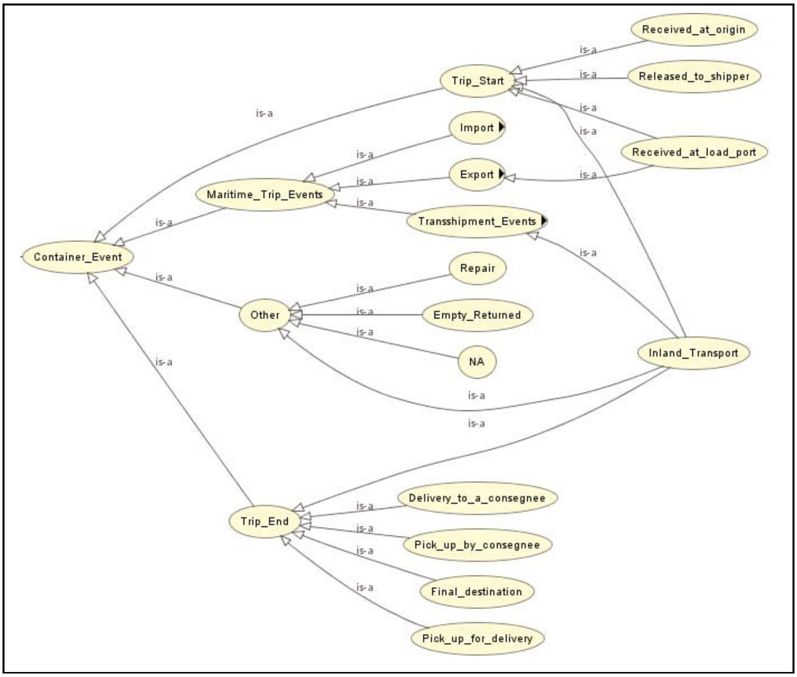


Fig. 4. Container Events

also recorded, specifically transshipment from one vessel to another. Each container event refers to several information dimensions, including the *Container status* (i.e., empty, full), the *Location* where this event took place (e.g., ports in intra-customs transportation, train stations and cities in inland transportation), and the corresponding *Time* (e.g., 14th September 2020 at 9:00 AM). Moreover, for events referring to transportation, a *Mean of Transportation* (e.g., vessel, truck) is involved (see Fig. 2, role *hasEvent* defined in *Maritime Trip Event*).

There is no standard reference for such events, and each shipping company adopts a different description. In ConTraffic an effort towards standardization of container events has been promoted, and the outcome has been formalized in MCO: 19 detailed events have been defined (Fig. 4), which are classified into four classes of top-level events: *Trip Start*, *Maritime Trip Event*, *Trip End*, and *Other*. Such events have a correspondence with the top-level shipping phases. Each top-level event is further classified to characterize the main phases of a shipping. In particular, for *Maritime Trip Event* we further distinguish among: 1) *Export* ; 2) *Transshipment*; 3) *Import*.

3.3 Leveraging Events to Define the Semantics of Trajectories

In MCO, container events play a fundamental role to formalize the semantics underlying container trajectories and trajectories for moving objects in general. In particular, we distinguish between *Event Sequence*, that is an ordered collection of

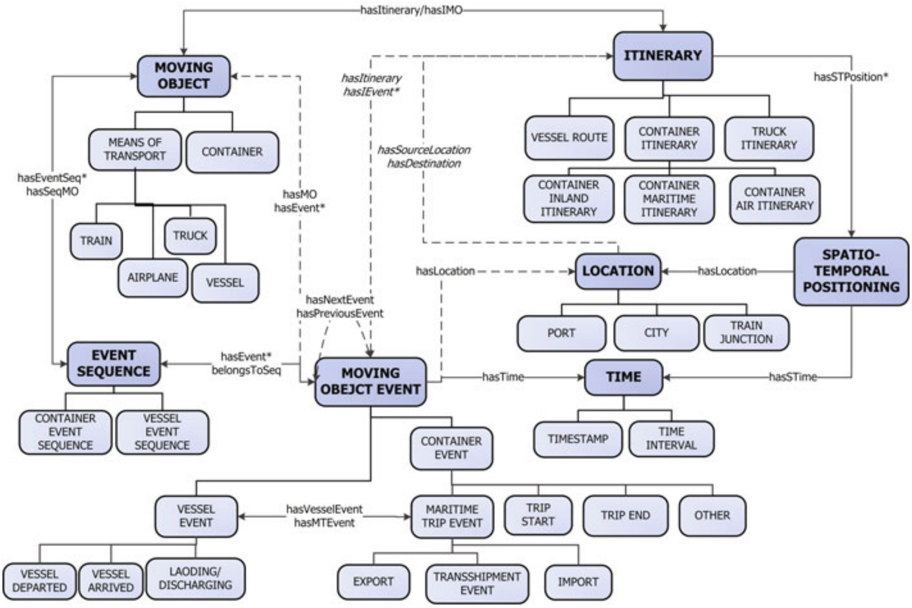


Fig. 5. MCO Events, Itineraries and Event Sequences

*Event*s that includes any activity done on a moving object, and *Itinerary*, which is the ordered sequence of events describing the movement of the object (see Fig. 5). Event sequences are intensionally defined by the transitive role *hasNextEvent* (and its inverse *hasPreviousEvent*) defined in *Event*; extensionally, they are formalized in the ontology by the concept *Event Sequence*.

For containers, an event sequence gives the lifeline of the container during the shipment, that can be further segmented into different shipment phases (see Fig. 3), contribution to satisfy the partWhole relationship that holds between *Shipping* and *Shipping Phase*. Among the events that define a *Container Event Sequence*, the event describing container movements contribute to the definition of the *Container Itinerary*. A container movement is implicitly represented by a location update in two subsequent events: whenever two events e_i and e_{i+1}, referring to locations l_k and l_m, respectively, are recorded in the MCO for container c, we may infer that c has moved from l_k to l_m. Corresponding shipment phases are defined to describe the transition.

The events involved in the maritime transportation of containers describe the trajectories between two import-export steps, which is modeled by *Maritime Container Itinerary*. Note that such a formalization may be viewed as a domain characterization of the STOPs and MOVEs model defined in [6,19], where *Container Events* characterize trajectories STOPs, while *Shipping Phases* define MOVEs.

A *Maritime Container Itinerary* is defined by *Maritime Trip Event* such as *Export* and *Import*, which give the start and the destination of the maritime trajectory of a container; *Transshipment Events* and its subclasses enable detailing of the maritime trajectory followed by a container during the intra-customs phase.

Whenever inland transportation events that characterize the start and the ending phases of a shipment are considered, the full container trajectory may be defined.

The concept *Maritime Container Itinerary* is fundamental for the definition of route based risk indicators for anti-dumping, which is a priority objective of ConTraffic. Indeed, the MCO formalisation of container itineraries, container events and shipment phases enable the application of different reasoning techniques to highlight inconsistencies in the behaviour of MCO actors. In the next section will discuss how the proposed formalisation is employed to perform axiomatic checking of suspicious movement patterns. However, other evaluations can be applied as well, like for example the recording and the comparison of mean time to accomplish certain handling time in ports.

Events that do not describe container transportation and that will not be used in the container itinerary definition describe, for example, deeds occurring within a port to prepare the container for the shipping or to complete it (e.g., loaded, deramped) and to accomplish the import-export procedures.

3.4 Vessels and Routes

Instances of *Means of Transport* represent trucks, trains, cargo ships and any other mean used for carrying goods in containers. We are particularly interested in modelling the behaviour of vessels, because most of the import-export of goods is performed by sea. In the MCO vessels are uniquely identified through their name and, where available, the International Maritime Organization (IMO) number. Every vessel transports hundred of containers in a single voyage, which is modeled by an instance of *Vessel Route* and gives an one-step itinerary from a starting port to a port of destination. The route followed by the vessel may not coincide with the maritime itinerary of a container, but partially overlap with it. Indeed, often the container is transshipped on another vessel to reach its port of destination, and at the same time other containers may be loaded on the vessel to reach the next port on its overall itinerary.

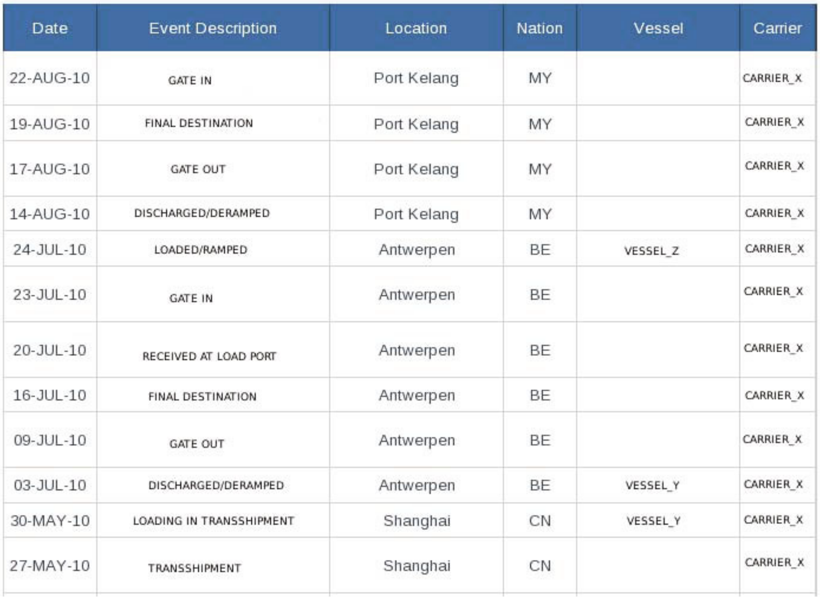
To model this situation, we define a class *Vessel Event* to represent the main events involving a vessel (e.g., Arrived at Port X, Loading, Discharging, Starting from Port Y). Such events map the subset of container events dealing with maritime transportation (i.e., subclasses of *Maritime Trip Event*). A sequence of events is defined through the transitive relationships *hasNextVesselEvent* and *hasPreviousVesselEvent*. Furthermore, we rely on the instances of *Vessel Event* to model the STOPs of a *Vessel Route*.

In particular, as described above, a *Transshipment Event* of a container involves two different vessels, therefore it is related with at least two different vessel events (a *Vessel Discharging* and a *Vessel Loading*). As we will see in Section 4, this implies also that, to check the consistency of a container itinerary that involves a transshipment, we have to compare it with two distinct *Vessel routes*.[2]

[2] In the generic case, in which the container has been transshipped several times, we have to consider multiple vessel trajectories. Let the number of transshipments done on the container be n; then, the vessel itineraries involved are $n + 1$.

3.5 Populating MCO

Data stored in the ConTraffic database are used to populate the MCO. In particular, we import the historical *sequences* of container events collected by the system. As illustrated in Fig. 6, each container history includes the dates when the events that compose it took place, their locations, and their description, including the container status and, when available, the names of the vessels used for transport.



Date	Event Description	Location	Nation	Vessel	Carrier
22-AUG-10	GATE IN	Port Kelang	MY		CARRIER_X
19-AUG-10	FINAL DESTINATION	Port Kelang	MY		CARRIER_X
17-AUG-10	GATE OUT	Port Kelang	MY		CARRIER_X
14-AUG-10	DISCHARGED/DERAMPED	Port Kelang	MY		CARRIER_X
24-JUL-10	LOADED/RAMPED	Antwerpen	BE	VESSEL_Z	CARRIER_X
23-JUL-10	GATE IN	Antwerpen	BE		CARRIER_X
20-JUL-10	RECEIVED AT LOAD PORT	Antwerpen	BE		CARRIER_X
16-JUL-10	FINAL DESTINATION	Antwerpen	BE		CARRIER_X
09-JUL-10	GATE OUT	Antwerpen	BE		CARRIER_X
03-JUL-10	DISCHARGED/DERAMPED	Antwerpen	BE	VESSEL_Y	CARRIER_X
30-MAY-10	LOADING IN TRANSSHIPMENT	Shanghai	CN	VESSEL_Y	CARRIER_X
27-MAY-10	TRANSSHIPMENT	Shanghai	CN		CARRIER_X

Fig. 6. An example of a container event sequence

In ConTraffic, 35,000 locations are currently mapped through their names and the names of the countries to which they belong. For most locations, also the United Nations Code for Trade and Transport Locations (UN/LOCODE)[3], assigned by the United Nations Economic Commission for Europe (UNECE), is stored. UN/LOCODE is a reference repository for commercial locations that is used by major shipping companies and currently maps over 76,300 locations worldwide.

Container events are *ordered* with respect to their temporal dimension and uploaded as instances of the appropriate *Container Event*. The relationships *hasNext* and *hasPrevious* and the corresponding instance of *Container Sequence* are filled in accordingly. The *Container Sequence* is segmented into different trips, defining the *Container Itinerarie*s. The *Vessel event*s corresponding to the *Maritime Trip Event*s are filled in, and the corresponding *Vessel Route*s

[3] http://www.unece.org/cefact/locode/locode_since1981.htm (accessed in December 2010).

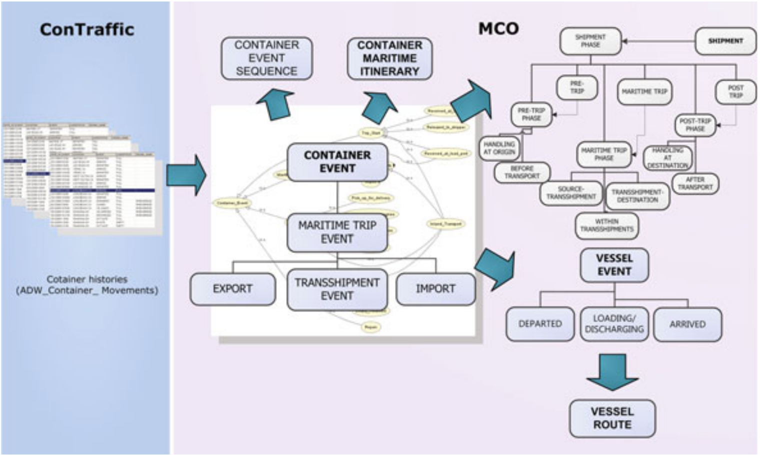


Fig. 7. MCO population

are generated. Finally, the corresponding relationships to fill in the details of *Shipping*s, are inserted.

4 SemRis to Formalize Suspicious Patterns

Once the semantic model to formalize the domain knowledge has been defined, we develop the SemRis to discover anomalous patterns. In this section, we use DL syntax [3] to identify two container itineraries that appear suspicious, because they involve extra cost/time in carrying out apparently unnecessary operations.

Specifically, we are interested in identifying two kinds of suspicious patterns:

***Loop pattern*:** The container shipped from its originating port P_1 is transshipped on another vessel that goes back to the originating port before reaching the destination port P_2 (cf. Fig. 8).

***Unnecessary_Trans pattern*:** The container from its original vessel (A) is transshipped on another vessel (B), in an intermediate port. From route reconstruction, vessel A also goes to the same destination. The container *appears* to be coming from the port of vessel B (cf. Fig. 9).

The corresponding SemRis are formalized as DL axioms, such that each axiom is a combination of logical operators that implicitly describes a class of objects. To ease their comprehension, in Fig. 10 we highlight the part of the MCO, opportunely expanded with respect to Fig. 2, which is involved in the axiom formalisation.

SemRI 1 is the axiom that formalizes the Loop pattern.

SemRI 1. *(Loop)* Given the MCO formalisation represented in Fig. 2 and Fig. 10, the DL axiom to describe the situation in which the container suspiciously comes back to the starting location P_1 before reaching the final one

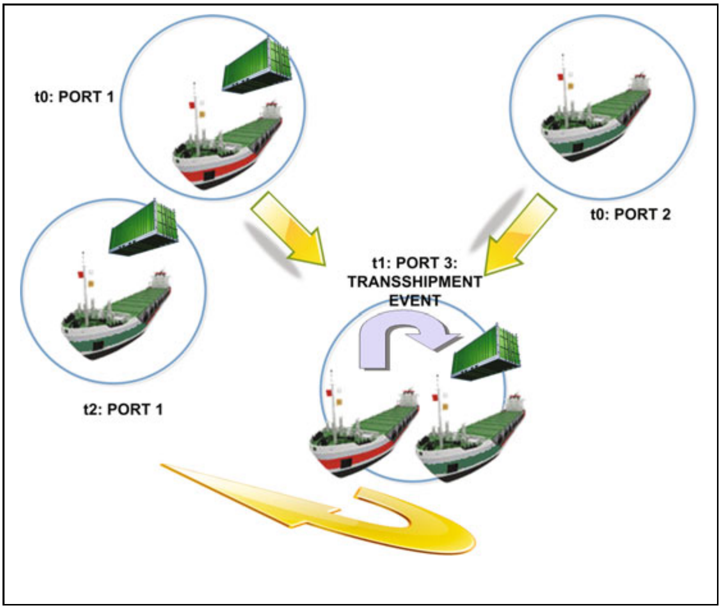


Fig. 8. Suspicious pattern Loop

P_2, as depicted in Fig. 8, is as follows:

$$\begin{aligned}\texttt{Loop} \equiv & \texttt{MaritimeContainerItinerary} \sqcap \exists\texttt{hasCISourcePort.P1}\sqcap\\ & \exists\texttt{hasCIDestinationPort.P2}\sqcap\\ & \exists\texttt{hasMaritimeTripEvent.(Transshipment\_Event}\sqcap\\ & \exists\texttt{hasLoadingVesselEvent.}\exists\texttt{hasNextVesselEvent.}\\ & (\exists\texttt{hasVEPort.P1} \sqcap \exists\texttt{hasNextVesselEvent.}\exists\texttt{hasVEPort.P2})) \qquad \square\end{aligned}$$

The core of Loop relies on the connection between the container events and the vessel events, formalized in the MCO by the role `hasLoadingVesselEvent`: such role links the description of the container itinerary to the route of the involved vessel. We have to observe that, since we are using DL syntax, we need to explicitly define a concept for each involved location (in the example, P_1 and P_2). This could be avoided using OWL syntax to formalize the axiom.

The following SemRI is a DL axiom that describes the Unnecessary_Trans pattern.

SemRI 2. *(Unnecessary_Trans)* Given the MCO formalisation represented in Fig. 2 and Fig. 10, the DL axiom describing the situation in which a container is sent to the port P by means of a transshipment, but the originating vessel goes to the same destination, as depicted in Fig. 9, is as follows:

$$\begin{aligned}\texttt{Unnecessary\_Trans} \equiv & \texttt{MaritimeContainerItinerary} \sqcap \exists\texttt{hasCIDestinationPort.P}\\ & \sqcap \exists\texttt{hasMaritimeEvent.(Transshipment\_Event}\sqcap\\ & \exists\texttt{hasDischargingVesselEvent.}\exists\texttt{hasNextVesselEvent.}\\ & \exists\texttt{hasVEPort.P}) \qquad \square\end{aligned}$$

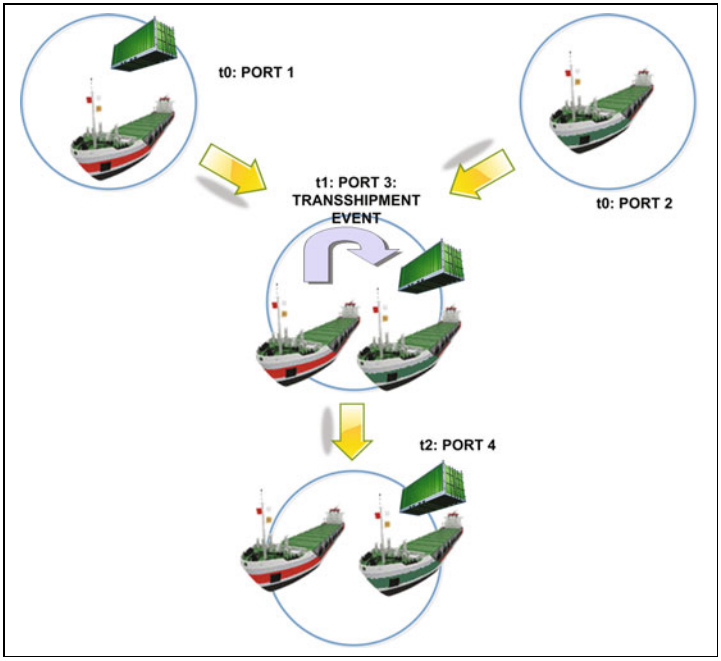


Fig. 9. Suspicious pattern Unnecessary_Trans

Also in this example, the main part of this axiom is represented by the connection between the container events and the vessel events: in this case, the role `hasDischargingVesselEvent`, that allows to pass from the description of the container itinerary to the one that brought it to the transshipment port. We have to point out that this axiom cannot be used without further elaboration criteria, because it matches all the ships that have the port P in their route after the transshipment. For the moment, a workaround to solve this problem is to elaborate further on the results obtained by the axiom by considering the dates of the first vessel arrival to port P: if the date is close to the container arrive, then the transshipment is unnecessary and the extracted pattern becomes suspicious.

5 MCO Potentialities, Issues and Future Developments

In the previous sections we have shown a description logic formalism that supports the discovery of suspicious patterns from maritime container trajectories stored as MCO instances. Following the approach we adopted in Section 4, such a formalisation may be extended to search other types of patterns in MCO: it is sufficient to define the corresponding DL axioms, then include them in a TBox directly processable by a reasoner such as Pellet [18] to evaluate them.

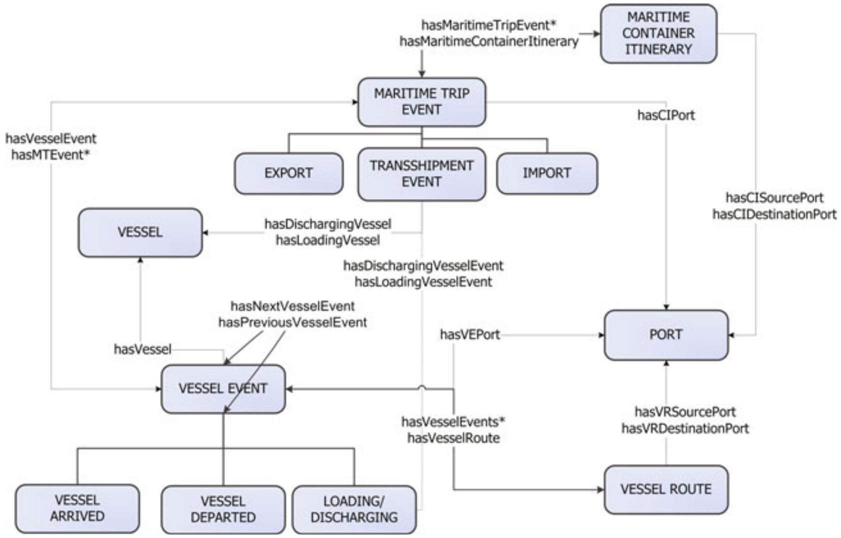


Fig. 10. Vessel and container events

With respect to other case studies addressed in the literature [4], the domain knowledge encoded by MCO and the reasoning necessary to discover anomaly are very complex. In particular, the detection of suspicious patterns requires multiple itineraries be analysed at the same time, i.e., container itineraries and vessel routes. Indeed, in Loop axiom the itinerary of one container is evaluated against one vessel itinerary, while in Unnecessary_Trans axiom the same itinerary is compared with two vessel routes. To partially reduce the inherent complexity of the domain, in MCO we introduced redundant relationships between concepts to avoid the specification of long paths in axioms. This simplifies their evaluation, but particular care is necessary when populating the corresponding roles in MCO. To avoid inconsistencies, this step can be formalised by triggering a set of rules written in Semantic Web Rule Language (SWRL, [15]), that extends OWL axioms to support Horn-like rules.

The discovery of patterns in itineraries is a step towards the semantic exploitation of trajectories for moving objects not only in the Maritime Surveillance area. Indeed, it is a problem of interest in GISScience, for the discovery of traffic patterns and way-finding for urban modelling, etc. Moreover, this work offers a formal study to develop an extended formalisation of search patterns, and it may be applied to discover anomalies in spatio-temporal sequences formalized as itineraries, for example intrusion detection in Secure areas.

The main advantage the proposed formalisation is the possibility to define axioms and properties in terms of high-level semantic concepts, abstracting away from different ways to describe the same events. Moreover, this approach enables to apply a DL reasoner to build an automatic system for the characterization of different itineraries in terms of the user's needs. In this respect, the approach is robust because the decidability of axiom evaluation is guaranteed by the DL formalism.

However, such an approach has the shortcoming, inherent to the DL approach and well known in the Formal Logic research field, of being hardly scalable on large sets of data. Since the search of suspicious patterns in MCO may involve the analysis of several thousands of records, the solution to this problem has a high dependency on the capability to manage large datasets. A potential solution to this problem is the usage of reasoning engines integrated by commercial database products, which are specifically designed to handle big knowledge bases formalized using standard semantic languages such as OWL and Resource Description Framework (RDF) and RDF Schema (RDFS) [9,10]. For example, Oracle Database Semantic Technologies [16] provides different tools (RDFS++, OWLSIF and OWLPrime) to infer new knowledge from semantic enabled repositories. However, such products currently lack of fundamental DL operations, such as union and intersection [16], which are necessary for axiom evaluation.

A viable approach to address the scalability issue might be the development of pre-processing procedures to reduce the size of the initial dataset, providing the DL reasoner with a smaller knowledge base in input. A similar approach has been adopted in [6], where an input dataset of touristic trajectories is first pre-processed with a set of data mining procedures to discover a bunch of data-mining patterns; then, such patterns are loaded in the knowledge base to reason on tourist behaviour.

Another drawback of DL is its limited expressive power, that sometime forces the development of artificial domain formalisations. For example, it is not possible to refer to variables in the axiom specification, therefore in Section 4 we had to specify each shipping port as a full semantic concept. However, weakening the decidability constraint, we can easily overcome this difficulty using formalisations such as OWL and SWRL that enable the use of variables and express equality comparison between instances (*owl:sameAs*).

Furthermore, another challenge of all semantic approaches in general is represented by the management of time: when we formalize a suspicious pattern with an axiom, in some cases we have to bound the analysis of vessel routes to avoid the inclusion of unrelated events that can result in false positive anomalous cases. This is the case of the Unnecessary_Trans axiom, that would match all the vessel routes going to port P after the transshipment of the given, even if they are no longer transporting it. Hence, we have to elaborate further on the results obtained by the axiom.

A proposal to handle time in ontologies is OWL Time Ontology [14], which is a vocabulary for expressing relations between instants, intervals, duration and date-time information. Unfortunately, a separate temporal reasoner is required to use it. To the best of our knowledge, so far there is no sound and complete integration of a standard reasoner and an OWL Time temporal reasoner capable of scaling to very large knowledge bases. For the moment, a solution is to elaborate further on the results obtained by the axiom by considering the date on which the vessel arrives at port P: if the date is close to the arrive of the container, then the extracted pattern becomes suspicious. In this respect, the application of *temporal granularities* [5] to abstract from detailed event dates to coarser (and somehow undetermined) timestamps may help to overcome this issue.

Furthermore, in our approach we also need to resolve a quality issue that affects the dataset. Indeed, ConTraffic system does not guarantee that all the events are collected, therefore *gaps* may exist in container histories. Such gaps, depending on their temporal coverage, complicate the interpretation of container sequences and their segmentation into trips, affecting the discover of trajectories and trades. Other inconsistencies arise because sometimes both actual and future events are collected. However, this problem is more relevant to the pre-processing phase than the successive ones.

In this sense, we have to point out that at the moment our approach is related only to complete itineraries: a possible extension of this research will be to integrate data mining technologies to manage incomplete itineraries. Moreover, since a peculiarity of such technologies is to discover implicit semantics, we can rely on them to manage unexpected patterns.

6 Conclusions

The ontology-driven enrichment of moving object itineraries seems to be promising in discovering anomalous itinerary patterns. In our work, we deal with suspicious patterns in maritime container itineraries, but the approach we propose may provide a contribution to the scientific community through a methodology to solve analogous problems for moving object itineraries in other application domains, such as traffic analysis, intrusion detection on the Web, etc. Indeed, the use of an ontological characterization of the system properties is flexible and the formalisation of anomalous patterns may be easily changed relying on the user's needs. Finally, as the ontology syntax is expressive, this approach can also be used to search for different types of itinerary patterns (e.g., to identify the most frequented itineraries).

Although the use of an ontology to describe the movement behaviour is not a "silver bullet", namely because at the moment there are scalability problems with large datasets, this approach has the advantage of enabling the specification of axioms and properties in terms of high-level semantic concepts, abstracting from the specific modelling adopted to represent the domain. Moreover, relying on such a formalisation we can develop an automatic system to characterize different itineraries patterns in terms of the user's needs.

In the next future, we plan to study the development of pre-processing procedures to reduce the size of the initial dataset. Moreover, we plan to investigate the employment of OWL and SWRL formalism in order to increase the expressiveness of our approach and we also need to resolve quality issues.

References

1. Alvares, L.O., Bogorny, V., de Macedo, J.A.F., Moelans, B., Spaccapietra, S.: Dynamic modeling of trajectory patterns using data mining and reverse engineering. In: Tutorials, Posters, Panels and Industrial Contributions at the 26th International Conference on Conceptual Modeling, ER 2007, vol. 83, pp. 149–154. Australian Computer Society, Inc., Darlinghurst (2007)
2. Alvares, L.O., Bogorny, V., Kuijpers, B., de Macelo, J.A.F., Moelans, B., Palma, A.T.: Towards semantic trajectory knowledge discovery. Technical report, Hasselt University (October 2007)
3. Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F. (eds.): The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press, Cambridge (2003)
4. Baglioni, M., de Macêdo, J.A.F., Renso, C., Trasarti, R., Wachowicz, M.: Towards semantic interpretation of movement behavior. In: 12th AGILE International Conference on Geographic Information Science, pp. 271–288 (2009)
5. Bettini, C., Jajodia, S., Wang, X.: Time Granularities in Databases, Data Mining, and Temporal Reasoning. Springer, Heidelberg (2000)
6. Bogorny, V., Heuser, C.A., Alvares, L.O.: A conceptual data model for trajectory data mining. In: Fabrikant, S.I., Reichenbacher, T., van Kreveld, M.J., Schlieder, C. (eds.) GIScience 2010. LNCS, vol. 6292. Springer, Heidelberg (2010)
7. Bogorny, V., Kuijpers, B., Alvares, L.O.: St-dmql: A semantic trajectory data mining query language. International Journal of Geographical Information Science 23(10), 1245–1276 (2009)
8. W3C Consortium. OWL (2004): The Web Ontology Language (2004)
9. W3C Consortium. RDF Vocabulary Description Language 1.0: RDF Schema (2004)
10. W3C Consortium. RDF/XML Syntax Specification, Revised (2004)
11. Donati, A.V., Kotsakis, E., Tsois, A., Rios, F., Zanzi, M., Varfis, A., Barbas, T., Perdigao, J.: Overview of the contraffic system. Technical report, JRC (November 2007)
12. International Organization for Standardization. Freight containers – Coding, identification and marking (1995)
13. Guc, B., May, M., Sayigin, Y., Koerner, C.: Semantic annotation of gps trajectories. In: 11th AGILE International Conference on Geographic Information Science (2008)
14. Hobbs, J.R., Pan, F.: Time Ontology in OWL (2006)
15. Horrocks, I., Patel-Schneider, P.F., Boley, H., Tabet, S., Grosofand, B., Dean, M.: SWRL: A semantic web rule language combining OWL and RuleML. W3C Member Submission (May 2004), http://www.w3.org/Submission/SWRL/ (last access on December 2008)
16. Oracle. Oracle Database Semantic Technologies Developer's Guide 11g Release 1 (11.1) (2009)
17. Orellana, D., Wachowicz, M., De Knegt, H., Ligtenberg, A., Bregt, A.: Uncovering patterns of suspension of movement (extended abstract). In: GIScience (2010)
18. Sirin, E., Parsia, B., Grau, B.C., Kalyanpur, A., Katz, Y.: Pellet: A practical OWL-DL reasoner. Journal of Web Semantics 5(2), 51–53 (2007)

19. Spaccapietra, S., Parent, C., Damiani, M.L., de Macedo, J.A., Porto, F., Vangenot, C.: A conceptual view on trajectories. Data & Knowledge Engineering 65(1), 126–146 (2008)
20. van Hage, W.R., de Vries, G., Malaisé, V., Schreiber, G., van Someren, M.: Spatial and semantic reasoning to recognize ship behavior (demo). In: ISWC (2009)
21. Zheni, D., Frihida, A., Ghezala, H.B., Claramunt, C.: A semantic approach for the modeling of trajectories in space and time. In: Proceedings of the ER 2009 Workshops (CoMoL, ETheCoM, FP-UML, MOST-ONISW, QoIS, RIGiM, SeCoGIS) on Advances in Conceptual Modeling - Challenging Perspectives, ER 2009, pp. 347–356. Springer, Heidelberg (2009)

Integration of Spatial Processing and Knowledge Processing through the Semantic Web Stack

Ashish Karmacharya[1,2], Christophe Cruz[2], Frank Boochs[2], and Franck Marzani[1]

[1] Institut i3mainz, am Fachbereich 1 - Geoinformatik und Vermessung
Fachhochschule Mainz, Holzstrasse 36, 55116 Mainz
{ashish,boochs}@geoinform.fh-mainz.de
[2] Laboratoire Le2i, UMR-5158 CNRS,
Dep. Informatique IUT Dijon, 7, Boulevard Docteur Petitjean
BP 17867, 21078 Dijon CEDEX, France
{christophe.cruz,franck.marzani}@u-bourgogne.fr

Abstract. This paper presents the integration process of spatial technologies and Semantic Web technologies and its associated tool. The result of this work is a spatial query and rule engine of spatial. To do so, existing ontology with spatial elements is adjusted in order to process the spatial knowledge through spatial technologies. This paper outlines the methods and the processes of these adjustments and how results are returned by our tool. The SWRL and the SPARQL language are extended for spatial purpose and the existing OWL ontology wine is used as an application example.

Keywords: Spatial processing, Knowledge processing, OWL Ontology, Rule language, Query language, SPARQL, SWRL.

1 Introduction

The Semantic Web is a set of technologies complementing the conventional Web tools proposed by Sir Tim Berners-Lee. It is seen as the most likely approach to reach the goal of semantic interoperability. The Semantic Web is envisaged as an extension to the existing web from a linked document repository into the platform where information is provided with the semantic allowing better cooperation between people and their machines. This is to be achieved by augmenting the existing layout information with semantic annotations that add descriptive terms to web content, with meaning of such terms being defined in ontologies [1]. Ontologies play crucial role in conceptualizing a domain and thus play an important role in enabling Web-based knowledge processing, sharing and reuse between applications.

This research attempts to contribute through including the functionalities of the spatial analysis within the Semantic Web framework. Moving beyond the semantic information, it has opened the chapter of inclusion of other form of information within the Semantic Web framework. It is important in the sense of the development of the technology itself. This work should at least provide a certain vision towards the direction the technology should take to integrate new forms of data. It discusses the direction in terms of spatial integration [3, 5].

C. Claramunt, S. Levashkin, and M. Bertolotto (Eds.): GeoS 2011, LNCS 6631, pp. 200–216, 2011.

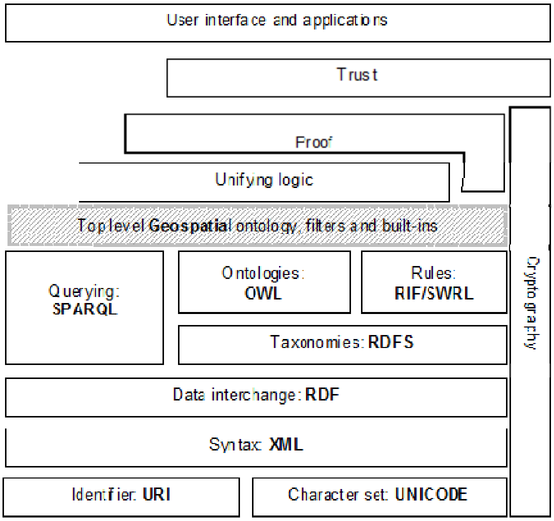


Fig. 1. Snapshot of the wine ontology adjusted with spatial component

The Semantic Web stack (e.g. fig. 1.) can be adjusted with a layer of that contains spatial information. The research proposes such an arrangement in the stack. A layer of spatial data mixing seamlessly with the semantic proposition in the layer Ontology through its OWL/RDF based syntax can be envisaged. This layer since uses the standard syntax of OWL/RDF can perform spatial queries through SPARQL or infer rules through standards as SWRL.

The integration process of the spatial technologies into the Semantic Web stack is undertaken by defining a new kinds of FILTERs for SPARQL queries and new kinds Built-ins for SWRL rules. These new FILTERS and Built-ins allow to process queries and rules with spatial data related to semantic data. The next chapter discusses this adjustment in the Semantic Web stack. Section 3 presents the top level ontology which enables the use of spatial technologies for any OWL ontology. Section 4 presents the translation engine which allows the translation of spatial queries and rules into standard queries and rules. Section 5 gives the complete process of extending an existing ontology in order to take into account the spatial technologies by using the famous wine ontology. Section 6 concludes this paper.

2 Background

The research project of Klein E. M. [2] in a certain degree follows the pattern of existing studies in geo-ontology research domain by focusing on the use of ontology for achieving data interoperability. It follows the pattern through defining the problems of data discovery in WFS (Web Feature Services) and the semantic differences in the data though the features associated to the data have same naming conventions. The problems follow even after the data discovery due to the nature of information that data represents is not explicitly stored. The research hence plans a mechanism of match making of different SDIs (Spatial Data Infrastructure) through the mediation of semantic rich domain ontology designed through the consultation of the experts. The Domain Ontology as it terms contains explicit information which capture the meanings of real world entities.

As with the case of this research, the research [2] utilizes the inference capabilities of the description logics in the ontology representation language of OWL-DL through inference rules. In addition, it uses a simple hydrological example to semantically annotate the data through the spatial rules. The SWRL representations of the rule are given:

```
Region(?x) ^ hasSlope(?x, Flat) → Lowland(?x)                        (1)
```

```
Lowland(?x) ^ River(?river) ^ adjacentTo(?x,?river) ^
hasAltitude(?x, ?xAlt) ^ hasAltitude(?river, ?riverAlt) ^
swrlb:subtract(?diffAlt, ?xAlt, ?riverAlt) ^                          (2)
swrlb:lessThan(4, ?diffAlt) → Floodplain(?x)
```

Region, Lowland, River and Floodplain are the concepts and hasSlope, adjacentTo and hasAltitude are the object properties in both feature type's ontology and domain ontology. The idea consists to semantically annotate the concept Floodplain with the rules. The first rule represented by equation 1 forms the lowland if the slope of a region is flat. There are many constraints of a region being lowland but the research uses this rule to demonstrate the usability. A Digital Elevation Model (DEM) is used in background which intersects the inferred information and the dataset is annotated as lowland. In short the object property hasSlope is intercepted and run through an algorithm which combines the DEM dataset to determine the flat slope. Regions inferring these flat slops are then annotated as lowland. Extending the rule to equation 2, it uses object property adjacentTo and built-ins of SWRL to annotate the floodplain. The object property adjacentTo again needs to run an algorithm in collaboration to the spatial dataset to provide the result. This result again infers with the other axioms in the knowledge base to enrich itself. The adjacentTo object property utilizes buffer operation to determine the objects close to it. However, the operation is hidden from the users and is executed inside the algorithm. This execution enriches the knowledge base which could be inferred through standard rule of SWRL. The execution of buffer or any spatial operations are carried out through the spatial operations of ArcGIS. The semantic annotation through these rules is carried out to enrich the Domain Ontology thus negating any short coming of explicit semantics in feature type's ontology.

The method of inferring the rules first through execution of spatial operations at database or application level and then enriching the knowledge base matches with the current research work. However, the implications in both researches are different. The approach that current research undertakes is to enhance the Semantic Web technologies through integrating spatial components into the technology. It differs significantly with the former research [2] as it was conducted to use semantic web tools and techniques to answer specific GIS problems. Hence, the scale of application of Semantic Web techniques is relatively low in the previous research [2]. In other hand, it could be seen that the spatial operations and functions are used implicitly through object properties like hasSlope or adjacentTo which are terms of natural language. This might give ambiguity to the interpretations of these terms. For example the term adjacentTo can have two or more meanings as rightly quoted in the thesis report. It can be near to each other either through touching or not touching. So, the utilization of spatial operation should be based on these factors. If the adjacent to means that the objects are touching then the spatial operation “Touch” could be directly used instead of Buffer which is more resource dependent.

Contrary to [2], this research has taken the works forward to address these concerns. Instead of using the commonly used terms, it uses the spatial operations and functions terminology standardized by OGC [4]. Standard terms are proposed to formulate rules rather than using domain based terms.

The equation 3 illustrated the adjustment of object property adjacentTo directly through SWRL rules through spatial built-ins, which means that the individuals of River which are in the Buffer of 50m of a individual of the concept Lowland possesses a relationship (ObjectProperty) named adjacentTo that link the Rivers and Lowlands.

```
River(?x)^Lowland(?y)^Buffer(?x, ?y, 50)
→adjacentTo(?x,?y)                                         (3)
```

Thus, it could be seen that there is much more flexibility concerning the definition of spatial rules through standard spatial built-ins proposed here. Besides the spatial built-ins for SWRL, this research adds on spatial built-ins to SPARQL, the query language of semantic web tools which is not explicitly researched before. However, before using these spatial built-ins in an existing ontology, it is first necessary to adjust this one with top level concepts. This integration process allows the linking of spatial data to ontologies. This ontology adjustment process is a generic process which allows the adjustment of any existing ontology in order to process spatial queries and rules on it.

3 The Top Level Ontology

This ontology serves as a foundation ontology to which objects can be instantiated during the identification process of spatial elements. The axioms are the building blocks of ontology and hence these axioms in the context of top level ontology of the application should be discussed to provide an overview of the system. The main axioms of this top level ontology are:

Semantic - spatial:Feature
Geometric - shp:Shape
Geometric Relationship - shp:hasShape
Spatial Relationship - sa:hasSpatialRelations
Spatial Database Relationship - doc:hasDBDetails

A shape has a definition in a spatial database. An individual has a shape and has spatial relationships with other individuals which have a shape.

The class axiom spatial:Feature represents the spatial objects. This class axiom is the generalized class of any objects with spatial definition. This class is further specialized into classes representing the different objects such as vin:Winery or vin:Region for instance regarding the example at the end of this paper. The spatial:Feature has to be specialized classes into subclasses. This abstract class cannot be instanced but only the individuals which belong to a subclass of spatial:Feature can have a spatial attribute.

The next important class axiom is shp:Shape which stores the local coordinates of the objects identified in the excavation site. This generalized class is specialized into shp:_3D and shp:_2D sub classes to represent the dimensions of the coordinates. Currently, an orthophoto is used to identify objects on a map and hence the 2D coordinates are returned of the objects. Semantics of objects in the knowledge base are defined through object property feat:objRel. But before that they need to relate to their spatial signature that is to their coordinates. This is managed through the specialized object property of shp:hasShape. As mentioned the coordinate of the object is derived through the digitization are stored as an individual of shp:_2D. This instance stores the coordinate of object. Once both the object and its coordinates are enriched, shp:hasShape provides a relationship between them. For instance, the concept win:Region as a subclass of spatial:Feature has the property shp:hasShape which can be a shp:_2D or shp:_3D.

The annotations to the database are carried out through assigning semantics to the annotations as assigning the relevant database and its relevant table in which the data is stored. It also provides the connection to the spatial column in which geometries of the objects are stored. An object property doc:hasDBDetails under general class doc:hasDocumentDetails provides these attributive connections. The three data properties to address the semantics of spatial annotation part of connecting to the MBRs are doc:dbName, doc:spColumn and doc:tableName, three specialized classes of doc:hasDBDetails.

The spatial functions and operations return geometries on their executions. It is hence important to have provision to store these returned geometries in the ontology. A generalized class sa:spatialOperation is introduced in the top level ontology. Every spatial operation under geoprocessing functions is then adjusted as its subclass. The class hierarchy of sa:spatialOperation reveals that the subclasses within it are the classes which need to represent returned geometries in some form.

The four spatial processing functions which are discussed here are Buffer, Union, Intersection and Difference. These spatial functions compute new spatial geometries. These new geometries are also stored in the spatial database in order to be computed by future spatial functions. As a solution, we definition four new classes called sa:sp_Buffer, sa:sp_Union, sa:sp_Intersection and sa:sp_Difference which are of specialized classes of sa:spatialOperation. The classes here are instantiated when the spatial operation of this category is executed. The result of execution is stored within the instantiated individual as the data property feat:localPlacement.

The functions under this category need to take a feature to execute them. The feature are objects within class feat:Feature. In order to maintain a relationship between the spatial operations representing classes under sa:spatialOperation and features under feat:Feature in the ontology an object property sa:hasSpatialRelations is added in the top level ontology. The specialized property relates the individuals under sa:spatialOperation and feat:Feature. For example for every instance in class sa:sp_Buffer (sub class ofsa:spatialOperation) be a property sa:hasBuffer (specialized object property of sa:hasSpatialRelations) which relates the sa:sp_Buffer class to the classes specializing feat:Feature. There are also four sa:hasSpatialRelations defined corresponding to each geoprocessing functions (sa:hasBuffer, sa:hasUnion, sa:hasIntersection, sa:hasDifference). Besides theses object properties, data properties to correspond the attributive nature of the relationships are also adjusted in the top level ontology. A generalized data property sa:hasSpatialAttribute is introduced in the top level ontology. Other attributive properties as sa:hasBufferDistance (denotes the buffer distance of the buffer) are specialized properties of it.

Table 1. The Spatial Processing Functions

Funtions	Concept	ObjectProperty	Execution Method
Buffer	sa:sp_Buffer	sa:hasBuffer(x,c)	$\mathrm{sa{:}\,sp_{Buffer}}$ $\sqsubseteq\ \exists \mathrm{sa{:}\,hasBuffer.\,feat{:}\,Feature}$ $\sqcap\ \mathrm{sa{:}\,hasBufferDistance.}\{\mathrm{c}\}$ C is of float value providing the buffer distance
Union	sa:sp_Union	sa:hasUnion(x,c)	$\mathrm{sa{:}\,sp_{Union}}$ $\sqsubseteq\ \exists \mathrm{sa{:}}^{2}\ \mathrm{hasUnion.\,feat{:}\,Feature}$ $\sqcap\ (\geq 2\ \mathrm{hasUnion})$
Intersection	sa:sp_Intersection	sa:hasIntersection(x,c)	$\mathrm{sa{:}\,sp\_Intersection}$ $\sqsubseteq\ \exists \mathrm{sa{:}\,hasIntersection.\,feat{:}\,Feature}$ $\sqcap\ (\geq 2\ \mathrm{hasIntersection})$
Difference	sa:sp_Difference	sa:hasDifference(x,c)	$\mathrm{sa{:}\,sp\_Intersection}$ $\sqsubseteq\ \exists \mathrm{sa{:}\,hasIntersection.\,feat{:}\,Feature}$ $\sqcap\ (\geq 2\ \mathrm{hasIntersection})$
Intersection	sa:sp_Intersection	sa:hasIntersection(x,c)	$\mathrm{sa{:}\,sp\_Intersection}$ $\sqsubseteq\ \exists \mathrm{sa{:}\,hasIntersection.\,feat{:}\,Feature}$ $\sqcap\ (\geq 2\ \mathrm{hasIntersection})$

Table 2. The Georelationtionship Functions

Functions	ObjectProperties	Characteristics
Disjoint	sa:hasDisjoint(x,y)	Symmetric
Touches	sa:hasTouch(x,y)	Symmetric
Within	sa:hasWithin(x,y)	Transitive
Overlaps	sa:hasOverlaps(x,y)	■
Equals	sa:hasEqual(x,y)	Symmetric, Transitive
Crosses	sa:hasCrosses(x,y)	Symmetric
Intersects	sa:hasIntersect(x,y)	Symmetric
Contains	sa:hasContain(x,y)	Transitive

These functions demonstrate the spatial relations between objects hence they are very straightforward when adjusting in ontology. They can be directly adjusted through object properties within the top level ontology. These functions are adjusted as specialized object properties of sa:hasSpatialRelations. The execution pattern of every function in this category is executed in similar. The table 2 illustrates the steps of every spatial function following OGC spatial operation standards but this research thesis utilizes four operations to demonstrate the argument. Those functions are Disjoint, Touch, Within and Overlap which are represented through sa:hasDisjoint, sa:hasTouch, sa:hasWithin and sa:hasOverlaps subsequently.

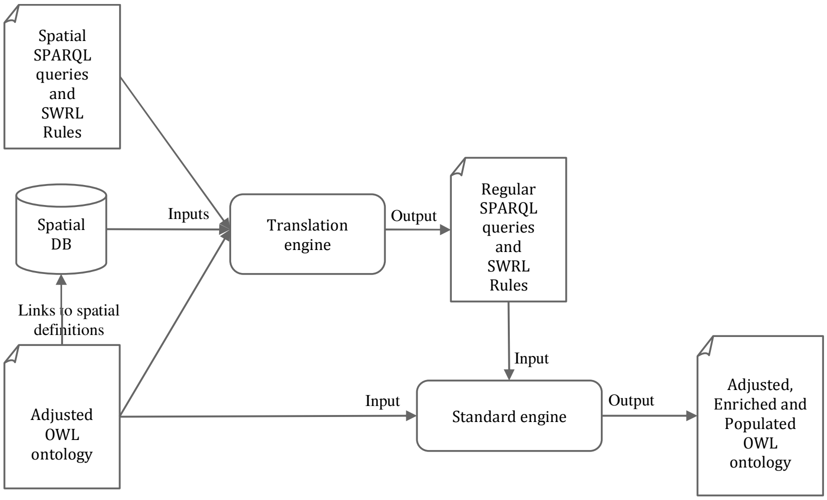

Spatial SPARQL queries and SWRL Rules
Spatial DB
Inputs
Translation engine
Output
Regular SPARQL queries and SWRL Rules
Links to spatial definitions
Adjusted OWL ontology
Input
Input
Standard engine
Output
Adjusted, Enriched and Populated OWL ontology

Fig. 2. The spatial processing of the translation Engine translating SPARQL queries and OWL rules

4 The Translation Engine

The translation engine allows the computation of spatial SPARQL queries and spatial SWRL rules. In both cases, the translation engine interprets the statements in order to parse the spatial components. Once the spatial components are parsed, they are computed through relevant spatial functions and operations by the translation engine through the operations provided at the database level. Then after, the spatial statements are translated to standard statements for the executions through their proper engine namely the SPARQL engine and the SWRL engine. Concerning the inference engine, the enrichment and the population of the ontology regarding the results of the inference process is possibly stored in the ontology.

The next sections presents in details the translation engine on more specifically the translation process of spatial SPARQL queries to regular queries. The following one presents the translation process of spatial SWRL rules to regular SWRL rules. These two processes have in common the use of SQL statements to query to the spatial database.

4.1 Spatial SPARQL Queries

The FILTER keyword in SPARQL queries is used to define spatial queries. A FILTER can be used to compare strings and derive results. The functions like regular expression which matches plain literal with no language tag can be used to match the lexical forms of other literals by using string comparison function. In addition, SPARQL FILTER uses the relational operators as = or > or < for the comparison and restrict the result. From this idea, the FILTER principle is extender in order to process georelationship functions.

Table 3. The spatial SPARQL syntax and its translation into SARQL syntax

Function	Spatial SPARQL Syntax	Translation
Buffer	SPATIAL_FILTER [buffer (?x, b, ?y)]	?x rel:hasBuffer ?y
	Result: Populated in the knowledge base as	?y rdfs:type sa:sp_buffer
	individuals of class sa:sp_Buffer.	?y sa:hasBufferDistance 200 000
Union	SPATIAL_FILTER [union (?x, ?y1,?y2)]	?x rdfs:type sa:sp_Union
	Result: Populated in the knowledge base as	?x sa:hasUnion ?y1
	individuals of class sa:sp_Union.	?x sa:hasUnion ?y2
Intersection	SPATIAL_FILTER [intersection (?x, ?y1,?y2)]	?x rdfs:type sa:sp_Intersection
	Result: Populated in the knowledge base as	?x sa:hasIntersection ?y1
	individuals of class sa:sp_Intersection.	?x sa:hasIntersection ?y2
Difference	SPATIAL_FILTER [difference (?x, ?y1,?y2)]	?x rdfs:type sa:sp_difference
	Result: Populated in the knowledge base as individuals of class sa:sp_Difference.	?x sa:hasDifference ?y1
		?x sa:hasDifference ?y2

4.1.1 Geoprocessing FILTER

The following example shows how to select Building which intersect the buffer of 200km of a River. In this example, the keyword FILTER is replaced by the keyword SPATIAL_FILTER in order to be processed by the translation engine

```
SELECT  ?name1 ?name2
WHERE
{
             ?feat1        feat:name      ?name1
             ?feat2        feat:name      ?name2
             ?feat1        rdfs:type      feat:River
             ?feat2        rdfs:type      feat:Building

             SPATIAL_FILTER [buffer (?x, 200 000,?feat1)]
             SPATIAL_FILTER [intersection (?y,?x,?feat2)]
}
```

This process is a selection process, and no inference process is engaged. Once the process is ended, the rule is translated to a standard given in the following example. It can be seen that the SPATIAL_FILTER is replace by standard RDF triples which. Any SPARQL engine is able to run this rule.

```
SELECT  ?name1 ?name2
WHERE
{
     ?feat1  feat:name                     ?name1
     ?feat2  feat:name                     ?name2
     ?feat1  rdfs:type                     feat:River
     ?feat2  rdfs:type                     feat:Building

     ?feat1  sa:hasBuffer                  ?x
     ?x      rdfs:type                     sa:sp_buffer
     ?x      sa:hasBufferDistance  200 000
```

```
      ?y      rdfs:type sa:sp_Intersection
      ?y      sa:hasIntersection ?x
      ?y      sa:hasIntersection ?feat2
}
```

The table 3 shows the translation of geoprocessing functions contained in SPATIAL_FILTER into standard triple component of a SPARQL query.

4.1.2 Georelationship FILTER

The following example shows how to select couples of features which are linked by a touch spatial relationship. In this example, the keyword FILTER is replaced by the keyword SPATIAL_FILTER in order to be processed by the translation engine. The name of features couples are selected with this restriction. The first feature has to be a feat:River which is of kind of feat:feature, and the second feature has to be a feat:Building which is also of kind of feat:feature. The SPATIAL_FILTER selects the couples which are touching spatially.

```
SELECT  ?name1 ?name2
WHERE
{
      ?feat1          feat:name       ?name1
      ?feat2          feat:name       ?name2

      ?feat1          rdfs:type       feat:River
      ?feat2          rdfs:type       feat:Building

      SPATIAL_FILTER [touches (?feat1, ?feat2)]
}
```

This process is a selection process, and no inference process is engaged. The aim of the translate engine consists to compute the touches spatial process of the Cartesian production between the features of the kind feat:River and feat:Building. In the case of a positive result, this new link is stored in the ontology between the couple of feature with the help of a sa:hasTouches relationship which is of the kind of sa:hasSpatialRelations. Once the process is ended, the rule is translated to a standard given in the following example. It can be seen that the SPATIAL_FILTER is replace by the triple "feat1 sa:touch ?feat2". Thus this rule can be processed by a standard SPARQL engine.

```
SELECT  ?name1 ?name2
WHERE{
      ?feat1          feat:name       ?name1
      ?feat2          feat:name       ?name2
      ?feat1          rdfs:type       feat:River
      ?feat2          rdfs:type       feat:Building          ?feat1
      sa:touch        ?feat2
}
```

The table 4 shows the translation of georelationship functions contained in SPATIAL_FILTER into standard triple component of a SPARQL query.

Table 4. The spatial SPARQL syntax and its translation into SARQL syntax

Functions	ObjectProperties	Characteristics
Disjoint	sa:hasDisjoint(x,y)	Symmetric
Touches	sa:hasTouch(x,y)	Symmetric
Within	sa:hasWithin(x,y)	Transitive
Overlaps	sa:hasOverlaps(x,y)	■
Equals	sa:hasEqual(x,y)	Symmetric, Transitive
Crosses	sa:hasCrosses(x,y)	Symmetric
Intersects	sa:hasIntersect(x,y)	Symmetric
Contains	sa:hasContain(x,y)	Transitive

4.1.3 Optimization

The translation engine is time consuming for large spatial database. In order to select the context of execution four options can be given to the SPATIAL_FILTER.

SPATIAL_FILTER_SELECT: No spatial operation is undertaken; the rule is translated without any spatial processing

SPATIAL_FILTER_PROCESS: Spatial operations are processed only for the couples of features which don't have this relationship. If this relation already exists, this one is not computed.

SPATIAL_FILTER_UPDATE: Spatial operations are processed only for the couples of features which have already this relationship in order to update these relationships.

SPATIAL_FILTER_ALL: This is the option by default which consists to compute all relationship for the Cartesian product in order to process it if it doesn't exist or in order or update it.

The following example shows that the selection of features which have the touches relationship is done with the option SPATIAL_FILTER_UPDATE.

```
SELECT  ?name1 ?name2
WHERE
{
          ?feat1         feat:name      ?name1
          ?feat2         feat:name      ?name2
             ?feat1             rdfs:type      feat:River
          ?feat2         rdfs:type      feat:Building

          SPATIAL_FILTER [touches (?feat1, ?feat2)]
          SPATIAL_FILTER_UPDATE
}
```

In addition the spatial filter can be combined by the following manner. It consists to insert news filters and to use the same variable. The following example consists to select building which contains a chimney in order to see if it touches a river. Moreover, no spatial processing is done, only the existing knowledge in the ontology is used to process this query.

```
SELECT  ?name1 ?name2
WHERE{
          ?feat1        feat:name       ?name1
          ?feat2        feat:name       ?name2
          ?feat1        rdfs:type       feat:River
          ?feat2        rdfs:type       feat:Building
          ?feat2        rdfs:type       feat:Chimney

          SPATIAL_FILTER [touches (?feat1, ?feat2)]
          SPATIAL_FILTER [touches (?feat2, ?feat3)]
          SPATIAL_FILTER_ SELECT
}
```

4.2 Inference Rules through SWRL

In an attempt to define the built-ins for SWRL, a list of eight built-ins was proposed during the research work. These eight built-ins reflect four geoprocessing functions and four georelationship functions that are discussed previously. The built-ins reflecting geoprocessing functions are built up in combinations with the spatial classes adjusted in the ontology and their relevant object properties. The built-ins for georelationship functions are in contrast are just object properties and using these object properties in collaboration to the spatial functions in database system.

4.2.1 Geoprocessing Built-Ins

The first set of built-ins is the built-ins for geoprocessing functions. They are functions returning geometries and adjusted in the ontology through feat:Feature sa:hasSpatialRelations sa:spatialOperation sequence. This class-property series is illustrated in table 5. The initial step consist the built-ins parsed to be processed by the translation engines. First the spatial built-ins are identified from the statement and parsed. Concurrently, the features on which these built-ins are applied are also identified.Then after, the SQL statementswith relevant spatial function on the relevantobjects of the featuresare executed at the database level. The results are then enriched in the knowledge base. Once, the knowledge base is enriched, the spatial built-ins are broken down into standard feat:Feature sa:hasSpatialRelations sa:spatialOperation sequence to generate the standard SWRL statement which is executed through standard inference engines.

Table 5. GeoProcessing built-ins

Functions	Class	Object Property	Data Property	Built-ins
Buffer	sa:sp_Buffer	sa:hasBuffer	sa:hasBufferDistance	Buffer(?x, b, ?y)
Union	sa:sp_Union	sa:hasUnion	-	Union(?x,?y1,y2)
Intersection	sa:sp_Intersection	sa:hasIntersection	-	Intersection(?x,?y1,y2)
Diffrence	sa:sp_Difference	sa:hasDifference	-	Difference(?x,?y1,y2)

The execution of every built-in can be elaborated through first running down the spatial operation and then translating the statements with spatial built-in into standard SWRL statements. Simplifying the explanations with an example of feat:Feature(?x) ^ Buffer(?x, b, ?y)

suggesting the use of built-in Buffer on objects within the specialized classes of feat:Feature with the buffer distance. This statement is elaborated first through running the SQL statement with the spatial function buffer on each objects of the class to which it meant to run. That is if the statement is related to buffering walls, then each instance of class feat:Wall is taken and buffered through the execution of the SQL statement. The SQL statement with spatial function Buffer would look like:

Table 6. The SQL statements executions of geoprocessing built-ins for the spatial enrichment

Built-ins	SQL Statements	Translated Built-ins	Built-ins
swrlbspatial:Buffer(?x, b, ?y)	SELECT Buffer(geom::Feature, bufferDistance) Result: Populated in the knowledge base as individuals of class sa:sp_Buffer.	sa:hasBuffer(?x,?y) ^ sa:p_Buffer(?y) ^ sa:hasBufferDistance(?y, b)	swrlbspatial:Buffer(?x, b, ?y)
swrlbspatial:Union(?x,?y1,?y2)	Select Union(geom::Feature1, geom::Feature2) Result: Populated in the knowledge base as individuals of class sa:sp_Union.	sa:sp_Union (?x) ^ sa:hasUnion(?x, ?y1) ^ sa:hasUnion(?x, ?y2)	swrlbspatial:Union(?x,?y1,?y2)
swrlbspatial:Intersection(?x,?y1,?y2)	Select Intersection(geom::Feature1, geom::Feature2) Result: Populated in the knowledge base as individuals of class sa:sp_Intersection.	sa:sp_Intersection(?x) ^ sa:hasIntersection(?x, ?y1) ^ sa:hasIntersection(?x, ?y2)	swrlbspatial:Intersection(?x,?y1,?y2)
swrlbspatial:Difference(?x,?y1,?y2)	Select Difference(geom::Feature1, geom::Feature2) Result: Populated in the knowledge base as individuals of class sa:sp_Difference.	sa:sp_Difference(?x) ^ sa:hasDifference(?x, ?y1) ^ sa:hasDifference(?x,?y2)	swrlbspatial:Difference(?x,?y1,?y2)

SELECT Buffer(geom::Feature, bufferDistance)

Here, the geom are the geometries of the objects within specialized classes of feat:Feature. The result of this execution is then enriched in the knowledge base. Primarily, the rows in result are geometries which indicate the buffers of each object with certain buffer distance. The class sa:sp_Buffer is instantiated with objects representing every row and storing the buffer geometry and the buffer distance within them. Then after, it is time to translate the statement with the spatial built-in into standard form of SWRL statement which would be

feat:Feature(?x)^sa:hasBuffer(?x,?y)^sa:sp_Buffer(?y)^ sa:hasBufferDistance(?y,b)

Thus, the statement converts the spatial built-in into feat:Feature sa:hasSpatialRelations sa:spatialOperation sequence of standard SWRL statement. The complete list of SQL execution, the result enrichment and statement translation process is illustrated in table 6.

The georelationship built-ins rely on object properties and more straight forward. The built-ins and their linkage to the object properties are presented in table 7.

Table 7. Georelationship Built-ins

Functions	Class	Object Property	Built-ins
Disjoint	-	sa:hasDisjoint	Disjoint(?x, ?y)
Touches	-	sa:hasTouch	Touches(?x, ?y)
Within	-	sa:hasWithin	Within(?x, ?y)
Overlaps	-	sa:hasOverlap	Overlaps(?x, ?y)

However, it is necessary to determine the nature of built-ins from the statement to determine what spatial operation needs to be performed at database level. These statements are hence parsed to identify the spatial built-ins from the statement. Then after, the SQL statement with related spatial operation is executed in the database level. The results are enriched against their specified object properties in the knowledge base. Now, the statements are ready to get executed. The spatial built-ins are broken down into feat:Feature sa:hasSpatialRelations feat:Feature sequence by the translation engine which is now a standard statement so can be executed.

Table 8. SQL statements executions of georelationship built-ins for the spatial enrichment

Built-ins	SQL Statements	Translated Built-ins
swrlbspatial:Disjoint(?x, ?y)	SELECT Feature2 FROM spTable WHERE Disjoint(geom::Feature1, geom::Feature2)	sa:hasDisjoint(?x, ?y)
swrlbspatial:Touches(?x, ?y)	SELECT Feature2 FROM spTable WHERE Touch(geom::Feature1, geom::Feature2)	sa:hasTouch(?x, ?y)
swrlbspatial:Within(?x, ?y)	SELECT Feature2 FROM spTable WHERE Within(geom::Feature1, geom::Feature2)	sa:hasWithin(?x, ?y)
swrlbspatial:Overlaps(?x, ?y)	SELECT Feature2 FROM spTable WHERE Overlap(geom::Feature1, geom::Feature2)	sa:hasOverlaps(?x, ?y)

It would be helpful to elaborate with an example of built-in

```
Feat:Feature(?x) ^ feat:Feature(?y)^ Touch(?x,?y)
```

It is a spatial operation to determine whether an object is touching another. Generally, the georelationship operations are binary operations and return Boolean values when is executed alone. However, when executed as a conditional parameter of the SQL statement, they yield results. That is if the statement

```
SELECT Touch(geom::Feature1, geom::Feature2)
```

is executed. It returns either true or false determining whether the geometry of feature1 touches geometry of feature2. But if the same operation is executed as

```
SELECT Feature2 FROM spTable WHERE Touch(geom::Feature1,
geom::Feature2)
```

then it returns all the feature2 which touches feature1. Here spTable is the table where the geometries of the features are stored in the database system and has been spatially annotated. The results derived through the execution of the statement with Touch operation is then enriched against sa:hasTouch object property of the specified feature. The last step is to break down the Touch(?x, ?y) built-in into feat:Feature sa:hasSpatialRelations feat:Feature sequence to get the SWRL statement executed. The breakdown of the spatialbuilt-in Touch(?x, ?y) is given as

```
  feat:Feature(?x) ^ hasTouch(?x, ?y) ^ feat:Feature(?y)
```

It is a standard SWRL statement which can again be inferred by inference engines. The complete list of SQL statement execution is illustrated in table 7.

5 The Wine Example

The famous wine ontology is used here to present the principle of spatial ontology adjustment which allows the computation of spatial data on any existing OWL ontology. The wine ontology is selected for several reasons. The wine ontology appears frequently in the literature as an example to define tutorials.

5.1 The Existing Ontology Adjustment

In order to adjust the exiting ontology, two main steps are essential. First, the top level ontology has to be imported into the existing ontology. In this manner, all the components of the spatial layer are available for the existing ontology, which are the annotation and tagging principles of documents and more specifically the spatial definitions. The second step consists to specialized specific concepts of the exiting ontology which have possibly spatial signatures. In the wine ontology, wine regions can be defined as spatial region or polygons in a GIS system. In addition, the wineries can be geolocalized as points in the same GIS system. Since the existing ontology is adjusted, the feed of the spatial database regarding the concepts respectively, wine region and wineries, of the ontology can be undertaken with the help of the individual already defined in the ontology. For instance, the individual vin:ClosDeVougeot, which is a French winery localized in Burgundy, is defined by the geolocalized point 47.174835,4.95544 in the WGS84 coordinate system.

The following figure shows the adjusted wine ontology. On the left side, the tree viewer represents the hierarchy of concept with the top level ontology and spatial:feature concept with the wine ontology specialized concept vin:Region and vin:Winery. All the other concept of the wine ontology can by spatially defined. On the right side, the list of the vin:Winery individuals is given. The individual vin:ClosDeVougeot appears in this list. This list is composed of 43 individuals and the list of vin:Region is composed of 36 individuals.

5.2 Spatial Querying Process

This section presents the benefit of spatial querying on spatial data composed of semantic definition. In the figure 5.16, the individual vin:CoteDOrRegion has a relationship has:adjacentRegion. This relationship defines a symmetric relationship between two regions. In the wine ontology, this information is not feed. Currently, it is no possible to select adjacent regions and regions which are around of 200km to each other. In the case of a spatial definition in a Spatial GIS, the following queries are possible. The first query select all the adjacent regions to vin:CoteDOrRegion. The second query select all the regions which are around of 200km to the region vin:CoteDOrRegion.

```
SELECT  ?adjacent
WHERE
{
     vin:CoteDOrRegion     rdfs:type      vin:Region
     ?adjacent             rdfs:type      vin:Region

        SPATIAL_FILTER [touches (vin:CoteDOrRegion,?adjacent)]
}

SELECT  ?region
WHERE
{
     vin:CoteDOrRegion     rdfs:type      vin:Region
     ?region        rdfs:type      vin:Region

        SPATIAL_FILTER [buffer(?buffer,200000,vin:CoteDOrRegion)]
        SPATIAL_FILTER [intersection (?res,?buffer,?region)]
}
```

The first examples of queries are related to the same kind of individuals, the same can be undertaken on different kind of individuals. For instance, no spatial relationships are defined between regions and wineries. With the adjustment of the ontology and the spatial definition of wine regions and wineries, now the following query can be undertaken easily. I would like to know all the wineries in a specific region.

```
SELECT  ?winery
WHERE
{
     vin:CoteDOrRegion     rdfs:type      vin:Region
     ?winery        rdfs:type      vin:Winery

     SPATIAL_FILTER [within (vin:CoteDOrRegion,?winery)]
}
```

If these relationships were defined in the ontology, then it would be possible to check the spatial consistency of the knowledge based. The individual vin:ClosDeVougeot is a winery located in vin:CoteDOrRegion. In the case of the definition of a symmetric relationship named vin:located between the concept vin:Region and the concept vin:Winery, the individual vin:ClosDeVougeot should be linked to the individual vin:CoteDOrRegion with the help of this relationship. The following query is able to validate this relationship from spatial point of view.

```
SELECT  *
WHERE
{
     vin:CoteDOrRegion     rdfs:type      vin:Region
     vin:ClosDeVougeot     rdfs:type      vin:Winery

        SPATIAL_FILTER [within (vin:CoteDOrRegion, vin:ClosDeVougeot)]
}
```

If the result is false, but the spatial data define and correct than the ontology is inconsistent. The overlap between the semantic links and the spatial data permits to check the consistency of the knowledge base in the case that the links were not generated from the spatial processing.

5.3 Spatial Inference Process

With the help of the SWRL rules, the enrichment of the ontology is now possible. The following simple example underlines this idea. The winery Clos de Vougeot vin:ClosDeVougeot which is a located in the region of Côte D'Or vin:CoteDOrRegion, and this region is actually a region located in France vin:FrenchRegion. Consequently, the winery Clos de Vougeot vin:ClosDeVougeot is located in France vin:FrenchRegion. The transitive relationship vin:hasSubRegion allows the definition of relationships between regions vin:Region.

This first SWRL rule enriches the ontology with vin:hasSubRegion relations between regions.

vin:Region(?x) ^ vin:Region(?y) ^ spatialswrlb:Within(?y, ?x) → vin:hasSubRegion(?x, ?y)

This second SWRL rule enriches the ontology with vin:isLocatedInRegion relations between wineries and regions.

```
vin:Region(?x) ^ vin:Region(?y) ^ vin:Winery(?z) ^
vin:hasSubRegion(?x, ?y) ^  vin:isLocatedInRegion (?z, ?x)
→ vin:isLocatedInRegion (?z, ?y)
```

This third SWRL rule does at the same time the first and the second rule by using spatial built-ins.

```
vin:Region(?x) ^ vin:Region(?y) ^ vin:Winery(?z) ^
swrlbspatial:Within(?y, ?x) ^ swrlbspatial:Within(?z, ?y)
→ vin:isLocatedInRegion (?z, ?y) ^ vin:hasSubRegion(?x, ?y)
```

After the execution of this third rule, new relationships vin:isLocatedInRegion and vin:hasSubRegion are created in the ontology in order to link . Consequently, the ontology is enriched with these new relationships.

6 Conclusion

This research attempts to highlight the possibilities to integrate spatial technology in semantic web framework. It moves beyond the scope of data interoperability while presenting the concept and makes efforts to utilize the potentiality in other areas of the Semantic Web technologies. The underlying technologies of knowledge processing provide to the semantic web the capabilities to process the semantics of the information through close collaboration with the machine. It makes not only the understanding of data easier for achieving interoperability among different data sources, but it also provides valuable knowledge which could enrich the knowledge base in order to equip it with new knowledge. This helps the users understand the data better. The underlying knowledge technology makes stand out among its contemporaries.

It is important to have standard terms for every built-in that will be developed to process spatial knowledge. With other built-ins in the tools standardized by W3C, the spatial built-ins should also get standardized by the consortium. In addition to W3C, OGC should also get involved in standardizing the built-ins. An effort in this direction should be carried out.

References

1. Horrocks, I., Pater-Schneider, P.F., McGuinness, D.L., Welty, C.A.: OWL: a Description Logic Based Ontology Language for the Semantic Web
2. Klien, E.M.: Semantic Annotation of Geographic Information. Essen: University of Muenster, thesis report (2008)
3. Karmacharya, A., Cruz, C., Boochs, F., Marzani, F.: Use of Geospatial Analyses for Semantic Reasoning. In: Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C. (eds.) KES 2010. LNCS, vol. 6276, pp. 576–586. Springer, Heidelberg (2010) ISBN 978-3-642-15386-0
4. OGC, The Open Geospatial Consortium, Inc.®, `http://www.opengeospatial.org/`
5. Karmacharya, A., Cruz, C., Boochs, F., Marzani, F.: ArchaeoKM: Managing Archaeological data through Archaeological Knowledge. In: Computer Applications and Quantitative Methods in Archeology - CAA 2010, Granada (Spain), April 6-9 (2010)
6. Bechhofer, S., Harmelen, F.v., Hendler, J., Horrocks, I., McGuinness, D.L., Patel-Schneider, P.F., et al.: OWL Web Ontology Language (February 10, 2004) (retrieved November 27, 2009), from W3C Recommendation: `http://www.w3.org/TR/owl-ref/`
7. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific AmericaN, 34–43 (May 2001)
8. Berry, J.K.: GIS Technology In Environmental Management: a Brief History, Trends and Probable Future. In: Soden, D.L., Steel, B.S. (eds.) Handbook of Global Environmental Policy and Administration, pp. 49–81. Decker, Marcel Inc., New York (1999)

Towards Heterogeneous Resources-Based Ambiguity Reduction of Sub-typed Geographic Named Entities

Mauro Gaio and Van Tien Nguyen

Laboratoire LIUPPA, BP-1155, 64013 PAU Université Cedex
{mauro.gaio,vantien.nguyen}@univ-pau.fr

Abstract. The aim of this work is to find sub-typed Geographic Named Entities from the analysis of relations between Place Names surrounded nominal group within a specific phrasal context in a set of textual documents. The paper presents a method involving natural language processing and heterogeneous resources like gazetteers, thesauri or ontologies. The work and the results focus a French language corpus. However, the uses of quite generic lexico-syntactic patterns in pre-selected phrasal context can be tuned for others languages.

Keywords: natural language processing, Named Entity categorization, verbal relations, lexico-syntactic patterns, finite-state transducers.

1 Introduction

Many applications such as automatic ontology creation, enrichment from text, information extraction, automatic indexation for digital libraries and question answering applications rely on different types of approaches. Currently, the first task consists in finding key terms (such as named entities and associated technical terms) in text of selected repository of documents and then use them as seeds for the next process. In automatic ontology enrichment these key terms are related to concepts in the target ontology.

Linguistic processing system using rule-based grammar and lexical resources may carry out this core task. Generally, this processing is combined with the recognition and classification of Named Entities [Nan98] and traditionally this process classifies named entities into persons, organizations and places. At the same time, research on analysis of geographic references is becoming a hot topic in the research area of information retrieval. So, "*. . . Finding geographical references in text is a very difficult problem and there have been many papers that deal with different aspects of this problem and describe complete systems such as Web-a-where, MetaCart, and STEWARD . . .*" [BLPD10].

However, in NER systems places entities are not classified in their specific sub-types and it is essentially due to the difficulty of the task [LL07]. In this paper, we focus on the identification and the geographical named entities subcategorization, gathered with two classes of additional lexical information. The

C. Claramunt, S. Levashkin, and M. Bertolotto (Eds.): GeoS 2011, LNCS 6631, pp. 217–234, 2011.

first class indicates a contextual geographic focusing and the second one indicates the sub-type e.g. “the EASTERN PART(1) of the Aspe VALLEY(2)”. The second sub-class has been already proposed in [MTV07]. By subcategorizing location entities, we propose a method to reduce ambiguities, retaining these key terms as seeds for the next step of a given process.

The proposed method involves lexico-syntactic patterns and external heterogeneous resources such as gazetteers, non-specific thesaurus, ontologies and event structures [RBH10] expressed in a finite-state description. While the use of such patterns essentially is not new in itself (for example Hearst patterns [HEA92] unlike most previous work, refining Hearst initial patterns or defining patterns within a very specified domain, in these work we investigate the uses of quite generic lexico-syntactic patterns in pre-selected phrasal context.

We are currently working on a repository of documents in French from the nineteenth century, devoted to the Pyrenees (especially travel stories). A travel story is a genre in which the author describes one or more travels, people encountered, emotions, things seen and heard. These documents contain numerous geographic references to typed geographic named entities of a defined area (French Pyrenees Mountains), all our examples hereafter are extracted from this repository. In a previous work [LES07] we have explored phrasal contexts where geographical point of view is predominant.

We propose, on the one hand, an automatic data processing sequence marking the contextual geographic focusing and the key term (syntactically represented by a nominal group) candidate to be a sub-type. This context will be fetch with a lexico-syntactic pattern. The goal is to mark, if it exists, the nominal group involved in such a context as a real sub-type or a “good” candidate to be a real sub-type. The choice depends on the existence of such key terms in external heterogeneous resources.

For limiting such a heavy task, our method proposes to checks if there is a sufficiently strong relationship between nominal group’s participation in a particular linguistic relation and its capacity to evoke a geographical sub-type. In a travel story genre the more interesting geographic phrasal contexts are represented by descriptions of persons’ movements or landscape perceptions. We propose a full-implemented automatic process in order to localize this potential geographic phrasal context. This context will be fetch with finite-state transducers embedded in a lexico-syntactic pattern. The last step is to use various external resources to validate or reduce the ambiguity of its geographical meaning.

2 Problems and Background

The traditional named entity recognition task is a well-known problem in the natural language processing (NLP) tasks and Information Extraction and Retrieving (IE & IR). Many systems have been developed, mainly on English, to recognize and categorize the proper names appearing in the texts [DM00] but any of them are able to classify places into specific sub-types such as RIVER, GLACIER, PEAK, MOUNTAIN... The identification of the geographic names is a well known much more complex task that simply recognizing place names or locations from others Named Entities. We are mainly interested by this category: locations and their intrinsic ambiguity as related in [VJW07], [LL07], [LEI04]. Our goal is to find an existing sub-type to reduce this intrinsic ambiguity. For example, *in Artouste lake*, or *in the peak of Artouste*, the place *Artouste* have a different semantics and different spatial representation according to the geographic type carried out by *lake* or *peak* key terms. In other words, in a task of the sub-type location geometry recovery with a known label, the called upon resources and the strategies of interrogation of these resources will be much more accurate. Hereafter in table 1 a part of a travel story extracted in our corpus from the book: "Ascension au pic de Néthou , Platon de Tchihatcheff, 80 pages, 1842".

Table 1. A paragraph from the corpus "Travel stories"

> *[...]Après avoir contemplé, avec une admiration mêlée d'effroi,* ***la charpente altière des Monts-Maudits****, nous songeâmes bientôt à descendre* ***sur le territoire aride de l'Aragon****. Le temps était menaçant : de légers brouillards parcouraient les hauteurs, et précédaient des nuages d'une teinte grisâtre, qui roulaient vers nous, venant* ***de l'ouest des Pyrénées****, un orage s'amoncelait : il ne tarda pas à éclater. Ayant renvoyé nos chevaux et payé le tribut accoutumé à la complaisance des carabineros (douaniers) espagnols, nos guides chargèrent nos provisions sur leurs épaules, et nous descendîmes, assez lestement,* ***vers le pied de la Maladetta****, laissant à notre* ***droite les roches calcaires de la Pèna-Blanca****. Arrivés au fond de* ***la vallée du Plan-des-Etangs****, qui est plus élevée que sa voisine,* ***la vallée latérale de l'hospice de Bagnères****, de 446 mètres, nous laissâmes derrière nous une cabane habitée pendant l'été par des bergers espagnols, pour remonter,* ***par un plan rocailleux, jusqu'au gouffre de Tourmon****, qui absorbe les eaux d'un torrent rapide, descendant* ***de la partie orientale du glacier de la Maladetta[...]****.*

As we can see in the example and in agreement with ([JON94], [FM03] and [LL07]) taken on its own place name can already be of different category such as: simple pure place names (*Bagnères, Maladetta, Tourmon...*) composed of only one lexeme; complex pure place names (*Monts-Maudits, Pèna-Blanca,...*) composed of several lexemes; slightly mixed place names containing link-words (*Bagnères de Bigorre,...*). In textual document all of these categories of place name can be combined in a syntactic relation with nominal groups being able to add a specific geographical meaning and finally build a complex and heterogeneous Geographic Named Entity (GNE) where explicit sub-types play an important role (*sur le territoire aride de l'Aragon, la partie orientale du glacier de la Maladetta,. . .*[1]).

The GNE could be potentially ambiguous in different part of its linguistic structure. Our global goal is to propose a whole method for reduce ambiguity. The method combines for each different ambiguities specific treatments. This paper focuses on the problem of sub-type ambiguity where we have defined three cases:

[1] On arid territory (zone) of Aragon, the eastern part of Glacier Maladetta.

Proposition 1
(1) *multi-referent when two different key terms existing in an ontology of reference are associated to the same Place Name.*
(2) *neo-referent when a the key term associated to at least one Place Name is not directly present in ontology of reference.*
(3) *lacked-referent when a the key term associated to at least one Place Name is not present in ontology of reference but exists in a non-specific thesaurus.*

For this last category of ambiguity it has been necessary to find a linguistic method to filter out local phrasal context suggesting a potential geographic meaning: a spatial pattern. A spatial pattern is an aspect of the identifiable speech by particular spatial characteristics. Related works have shown that these characteristics result in linguistic aspects. For instance, in [TT97] authors define, study and categorise three patterns for description, the description of way, by course of the glance, and description in over flight. In agreement with [LES07], we have retained four spatial patterns: itinerary, local description, points of, places comparison. For each of these spatial patterns the main bootstrapping linguistic mark is the use of a specific verb category.

Thus, the study of our corpus relating to travel stories showed that whenever a place name is used it could be associated with a nominal group evoking a kind of more or less sharped spatial focusing called therefore indirection. More over if the place name associated or not with an indirection is evoked in a sentence containing a verb of movement or a verb of perception, then the nominal group between the verb and the place name frequently has a geographical connotation (in our corpus approximately 50% of the terms result from an ontology of topographic domain). Thus, we put forth the assumption that such event structures [RBH10] are interesting discriminating indicators for making a first selection of nominal groups used for their geographical mining.

In this paper, due to our corpus, we have particularly explored the two first patterns:

Proposition 2
(a) *The pattern itinerary corresponds to a description of a set of the author's movements from place to place, related to a journey.*
(b) *The pattern local description corresponds to a description of a restricted place, the speaker being in this place. This pattern is appropriate for a description made without movement on the part of the author.*

Both patterns enable us to apply filters bootstrapped by verbs. To put it in a nutshell, the pattern itinerary is characterized by verbs of displacement as defined in [M.08] and the pattern local description is characterized by verbs of perception and verbs of state. A description calls upon the five senses: sight, hearing, touch, taste and sense of smell. To evoke a feeling, the authors use above all the verbs of perception such as to see, to hear, to touch, to taste and to feel. Due to the specificity of our corpus of travel stories we have restricted to two categories of verbs: verbs of displacement and verbs of perception and we are looking to a particular semantic relation involving a verb of displacement or a verb of perception with or without a preposition in relation with or without one or more nominal groups. These observations also collaborate research done in displacements reference in language such as that of

We essentially uphold the criteria of a verb's aspectual polarity that was introduced by Boons [BOO87] and further developed in [LAU91]. Thus, we model the particular semantic relation in the text by taking account of displacement verbs that are necessarily associated with Place Names, and that are optionally associated with spatial clauses.

In the model we propose here, the triplet is discriminating to bring out the spatial meaning of certain polysemic verbs (*to leave someone* is of little interest to us, whereas *to leave Pau* attracts our attention). The same is true for *to get out of a tough spot vs to get out of Pau.* This means that the pattern to extract the particular semantic relation defers if the verbs involved is initial (*to leave*), median (*to cross*) or final (*to arrive*). These notions meet the LRV[2] thesis of Sablayrolles [AS95] who also studied the motion verbs.

3 Method and Implementation

In order to reduce different levels of ambiguity carried in different parts of GEN we use a methodology combining various lexico-syntactic patterns [HEA92], [MZB04], [MFP09], in a process taking into account phrasal context.

The core of our method is based on a lexico-syntactic pattern called from hereafter *VPT*. It's a triplet which is composed of the following elements : the first one being the verb of displacement, or verb of perception, the second one being preposition, and last one toponym. In some cases, a preposition can lack in the triplet VPT. In each triplet, a `toponym` is represented by the following non-ordered items list, `[sub-type candidate*, indirection?, place name]` where:

- `?` for specifying that there must not be more than one occurrence —the item is optional— ;
- `*` for specifying that any number (zero or more) of occurrences is allowed —the content of each occurrence may be different and are generally disconnected and the item is optional—;

Some of VPT' different instances are summarised in Table 2 and the fully implemented chain including the VPT' principles is illustrated in the figure 1. In this figure (a) represents major steps of our processing sequence; (b) explains the output of each step with the input sentence: *Nous songeâmes bientôt à descendre sur le territoire aride de l'Aragon*, which comes from the paragraph in table 1. Our chain is designed for processing a corpus in french, but it can be tuned for texts in other language. This adaptation will be discussed later. In the figure 1, to make paper more understandable to non-french speakers the input and outputs in each step have been translated in English.

[2] Lieu de Référence Verbal (Verbal Reference Location).

Table 2. Some examples of the triplet VPT

Verb	Preposition	Toponym		
		Sub-type candidate	Indirection	Place name
arriver (arrive)	à (to)	ville (city)	au sud de (in the south of)	Pau
aller (go)	dans (into)	vallée (valley)	au nord de (in the north of)	Paris
venir (come)	de (from)	rivière (river)	au centre de (in the centre of)	Aragon
voir (see)	vers (toward)	village (village)	à coté de (near from)	Azun

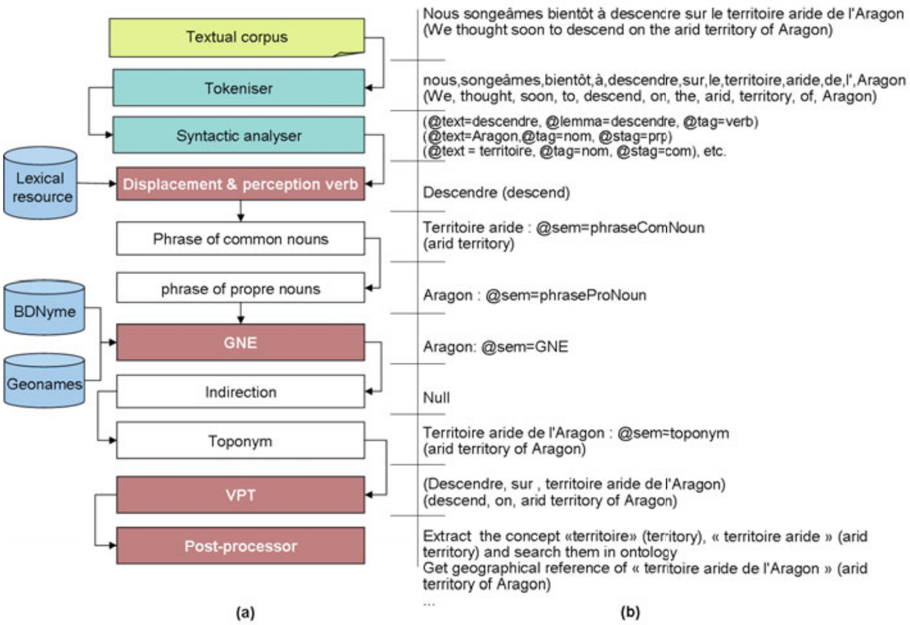

Textual corpus
Tokeniser
Syntactic analyser
Lexical resource
Displacement & perception verb
Phrase of common nouns
phrase of propre nouns
BDNyme
Geonames
GNE
Indirection
Toponym
VPT
Post-processor
(a)
Nous songeâmes bientôt à descencre sur le territoire aride de l'Aragon
(We thought soon to descend on the arid territory of Aragon)
nous,songeâmes,bientôt,à,descendre,sur,le,territoire,aride,de,l',Aragon
(We, thought, soon, to, descend, on, the, arid, territory, of, Aragon)
(@text=descendre, @lemma=descendre, @tag=verb)
(@text=Aragon,@tag=nom, @stag=prp)
(@text = territoire, @tag=nom, @stag=com), etc.
Descendre (descend)
Territoire aride : @sem=phraseComNoun
(arid territory)
Aragon : @sem=phraseProNoun
Aragon: @sem=GNE
Null
Territoire aride de l'Aragon : @sem=toponym
(arid territory of Aragon)
(Descendre, sur , territoire aride de l'Aragon)
(descend, on, arid territory of Aragon)
Extract the concept «territoire» (territory), « territoire aride » (arid territory) and search them in ontology
Get geographical reference of « territoire aride de l'Aragon » (arid territory of Aragon)
...
(b)

Fig. 1. Our processing main steps

3.1 Marking Verbs of Movement|Perception

Firstly, the text is tokenized before being processed by a syntactic analyser (i.e TreeTagger[3]) which associates each token to a grammatical category (i.e. verb, noun, preposition, etc.). Then, thanks to our lexical resource, verbs of movement and verbs of perception are marked. In accordance with the retained concept of aspectual polarity, the verbs of movement are also marker as: "initial verbs"

[3] TreeTagger is a language independent part-of-speech tagger. It was developed by Helmut Schmid in the TC project http://www.ims.uni-stuttgart.de/projekte/tc/ (at the Institute for Computational Linguistics of the University of Stuttgart).

or of initial polarity, for verbs like *quitter, partir, sortir, s'échapper, s'éloigner, etc*[4]. "Final verbs" of final polarity, for verbs like *arriver à, atteindre, entrer dans, regagner*[5]. "Median verbs" or of median polarity, for verbs like *traverser, descendre, franchir, parcourir, passer par, se déplacer dans, etc*[6].

An analysis of verbs of perception enabled us to conclude that all these verbs are transitive verbs, and thus they are never followed by a preposition i.e. all the verbs of perception has a sub-behavior of median verbs of movement (median polarity). Then, we have added to our resource a lexicon of about fifty verbs of perception.

3.2 Marking Toponyms

Basing on the output of the syntactic analyser, words or group of words are marked as *common nouns*, or as *proper nouns.* A single common noun could be, *vallée, village, territoire, etc*) and a complex one could be, *territoire aride, marché d'intérêt régional, etc.* Recursively, we separate the adjective(s) from the noun. This is done with rules expressed in a DCG (Definite Clause Grammar) formalism. This formalism allows to implement context-free grammar. In our case it consists of a set of rules to replace a sequence of speech (noun, adjective, verb, etc.) by a new unique identifier (noun phrase, verb phrase, etc.). Our rules marking the words or group of words as common nouns, presented in table 3, are expressed in Prolog[7]. In this table, line 3 shows how if a sequence of tokens contains an adjective which is located before a common noun (or a group of common noun, recursively), all the sequence will be represented as a common nouns. This kind of marked sequence will be retained to be candidate for sub-typing the place name.

Table 3. DCG marking the phrase of common noun

```
 1 root(commonNoun:X) --> group(X).
 2 %case 1 : ex : belle ville
 3 group(adjectif:A..nom:N) --> adjectif(A), group(N).
 4 %case 2 :ex : territoire aride
 5 group(nom:N..adjectif:A) --> commonNoun(N), adjectif(A).
 6 %case 3 :hotel de ville (recursively)
 7 group(nom1:N1..nom2:N2) --> commonNoun(N1), %hotel
 8         (ls_token('de');ls_token('d\'');ls_token('des')), %de
 9         group(N2). %ville
10 %cas4 : territoire
11 group(X) --> commonNoun(X).
12 commonNoun(adjectif:' '..nom:lemma:X) --> ls_token(_,
13 lemma:X..stag:com, token).
14  adjectif(A) --> A@tag:adj.
```

Similarly the phrase of proper nouns can be a single proper noun (i.e Aragon, Pau, etc) or a group of proper nouns connecting by some others words (*I.e.*

[4] To quit, to leave, to go out, to escape, to get away, etc.

[5] To arrive, to reach, to get in, to go back, etc.

[6] To cross, to go down, to overcome, to cover, to go by, to move in, etc.

[7] Prolog is a general purpose logic programming language.

Table 4. 3 among 14 DCG rules for marking the phrase of proper noun

```
root(lemma:X) --> group(X).
%ex1 : Mont de Marsant (recursively)
group(X) --> N1@stag:pro, ls_token('de'), group(N2),
{string_concat(N1, ' de ', S)},
{string_concat(S, N2, X)}.
%ex2: Saint Jean
group(X) --> N1@stag:pro, group(N2),
      {string_concat(N1, ' ', S)},
      {string_concat(S, N2, X)}.
%ex3 : Aragon
group(X) --> X@stag:pro.
...
```

Mont-de-Marsant, Saint-Jean-Pied de Port, etc). The DCG rules marking theses phrases are presented in table 4.

Gazetteers (BNNyme, Geonames, etc.) are used to validate phrases of proper nouns as existing Place Names, after each validation the phrase of proper noun is marked as a GNE. In next step, the indirection (i.e. au sud de, au centre de, etc) will be marked thanks to a specific lexical resource. Finally the toponym is defined as a composition of the elements marked in previous steps : the phrase of common nouns, the indirection, the GNE. This is realized thanks to a set of DCG rules (table 5). Note that we categorize toponyms into two main classes : *absolute and relative.* An absolute one is illustrated by an example in lines 1 to 8, and a relative one in lines 10 to 19.

Table 5. DCG for marking toponyms

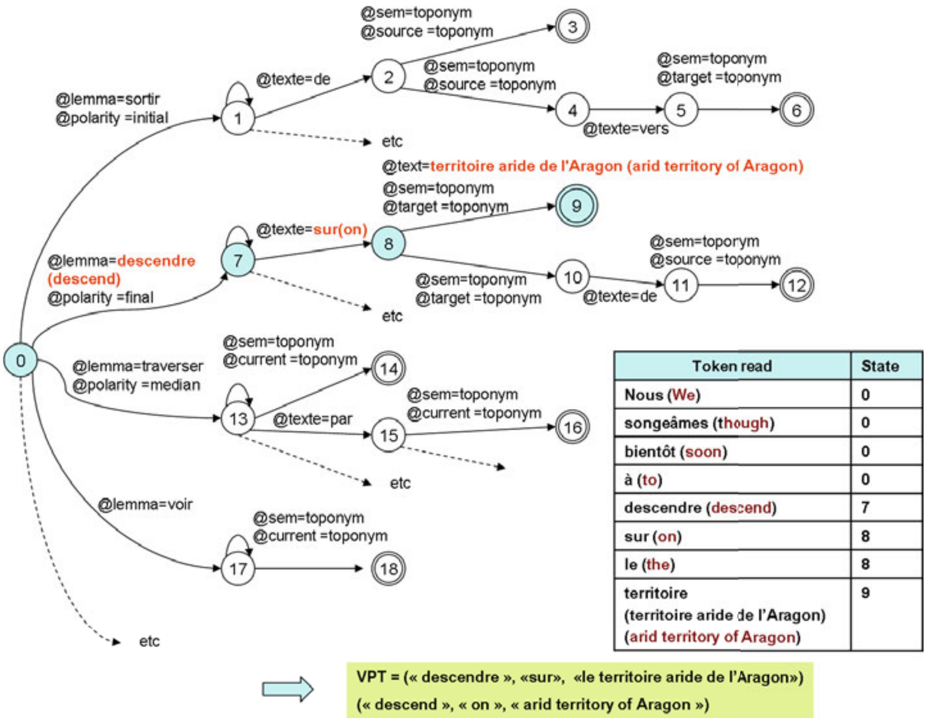
```
 1 Absolute toponym : territoire aride de l'Aragon (arid territory of
       Aragon)
 2 toponym(esa:X..type:a) --> esa1(X).
 3 Define absolute toponym
 4 esa1(subType:X..placeName:Y) --> subType(X), %territoire aride
 5    de,    %de
 6    placeName(Y). %Aragon
 7 subType(X) --> ls_token(_, X, commonNoun). %territoire aride
 8 placeName(X) --> ls_token(_, X, placeName). %Aragon
 9
10 Relative toponym : territoire aride au sud de la ville de Pau (arid
       territory in the south of Pau city)
11 toponym(esr:X..type:r) --> esr1(X);
12 Define the relative toponym
13 esr1(subType:X..indirection:Y..esa:Z) -->
14    subType(X), %territoire aride
15    indirection(Y), %au sud de
16    article, %la
17    esa(Z). %ville de Pau
18 esa(Z) --> esa1(Z); esa2(Z).
19 indirection(X) --> ls_token(_, lemma:X, indirection).
20 ...
```

Next task marks the linguistic structure VPT: Verb of movement (Vmov) or Verb of perception (Vperc), preposition and toponym.

3.3 Extracting the Structure VPT (Vmov|Vperc, Preposition, Toponym)

In agreement with [Lou08], this is done by the transducers. This section presents the analysis obtained thanks to these transducers. Two sentences, each one been a membership of a spatial pattern as previously presented in proposition 2.

Spatial pattern: itinerary. Verb of movement is the core of this pattern. Transducers are illustrated in figure 2. The processing of the pattern *itinerary* is explained hrough the sentence already given above. Tokens and transitions are given in the table as well as different state transitions of the transducer.



Token read	State
Nous (We)	0
songeâmes (though)	0
bientôt (soon)	0
à (to)	0
descendre (descend)	7
sur (on)	8
le (the)	8
territoire (territoire aride de l'Aragon) (arid territory of Aragon)	9

VPT = (« descendre », «sur», «le territoire aride de l'Aragon»)
(« descend », « on », « arid territory of Aragon »)

Fig. 2. Examples of transducers involving verbs of movement like: *sortir, arriver, descendre* and by extension a verb of perception like *voir*

The principle is quite simple, the transducer starts in state 0, and the text is processed token by token. Depending on the semantic of each token, the transducer passes into a new state or not. For example, there is no state changes when tokens "nous", "songeâmes", "bientôt", "à" are read. But when the token "descendre" (go down to) is read its semantic mark *verb of movement* allows the transducer to change its state from 0 to 7. So the first element of the triplet VPT (i.e. the verb "descendre") is marked. When the token containing the preposition "sur" (on) is reached the transducer passes into state 8. Finally, the toponym is marked when the token containing "territoire" (territory) is reached. In fact, this token belongs to a group of words (the toponym "territoire aride of Aragon") whose semantic mark GNE allows the transducer to pass into the state 9. So the toponym "territoire aride of Aragon" is marked. The transducer pass in its final state and returns the structure VPT (“descendre”,”sur“, “territoire aride of Aragon”) figure 3.

Spatial pattern: local description. This pattern may be characterised by a verb of perception. Consider an example sentence “J’ai vu la vallée au sud du village d’Azun” (I saw the valley in the south of Azun village). The figure 4 presents the state transition of our transducer for marking the triplet VPT. When the transducer passes in a final state: the structure VPT (“voir”,” “, "vallée au sud du village d’Azun”) is obtained.

The figure 5 represents the output exported by our processing chain for input as the paragraph in the table 1.

Thus, thanks to the transducers, we are able to analyze both types of verbs: verbs of movement and verbs of perception. In next step, the structure VPT will be useful for helping to reduce ambiguity in the GNE.

3.4 How the Structure Extracted is Useful in the Three Cases of Ambiguities?

Our previous example, "*nous songeâmes bientôt à descendre sur le* ***territoire aride de l’Aragon***" illustrates a lacked-referent case and how the method could not fully disambiguate the sub-type:

"le territoire aride de l’Aragon" has been marked as a toponym, but, neither key term territoire aride nor key term territoire are present in the domain-specific ontology of reference [8]. But this toponym is involved in the structure VPT, so the nominal group "territoire aride" (arid territory) could be considered as a “good” candidate to be geographic sub-type. So, the second external resource (the generic thesaurus RAMEAU) is queried and this time the concept *territoire* is found. Unfortunately its relations with others concepts do not permit to fully validated the GNE as a toponym.

Table 6 presents examples of key terms extracted in our corpus of travel stories not present in the domain-specific ontology but found in the generic thesaurus RAMEAU. The table also illustrate how conceptual relations can add or reduce ambiguities.

3.5 Can Our Method Be Reused to Process Other Languages ?

The processing schema figure 1(a) is designed to be natural language independent in the present work, due to the corpus, the process as been fully tested for French. For it use with another language, a tuned phase is necessary for

[8] In this work the domain-specific ontology has been established in collaboration with the COGIT a research group of IGN ([AM10]).

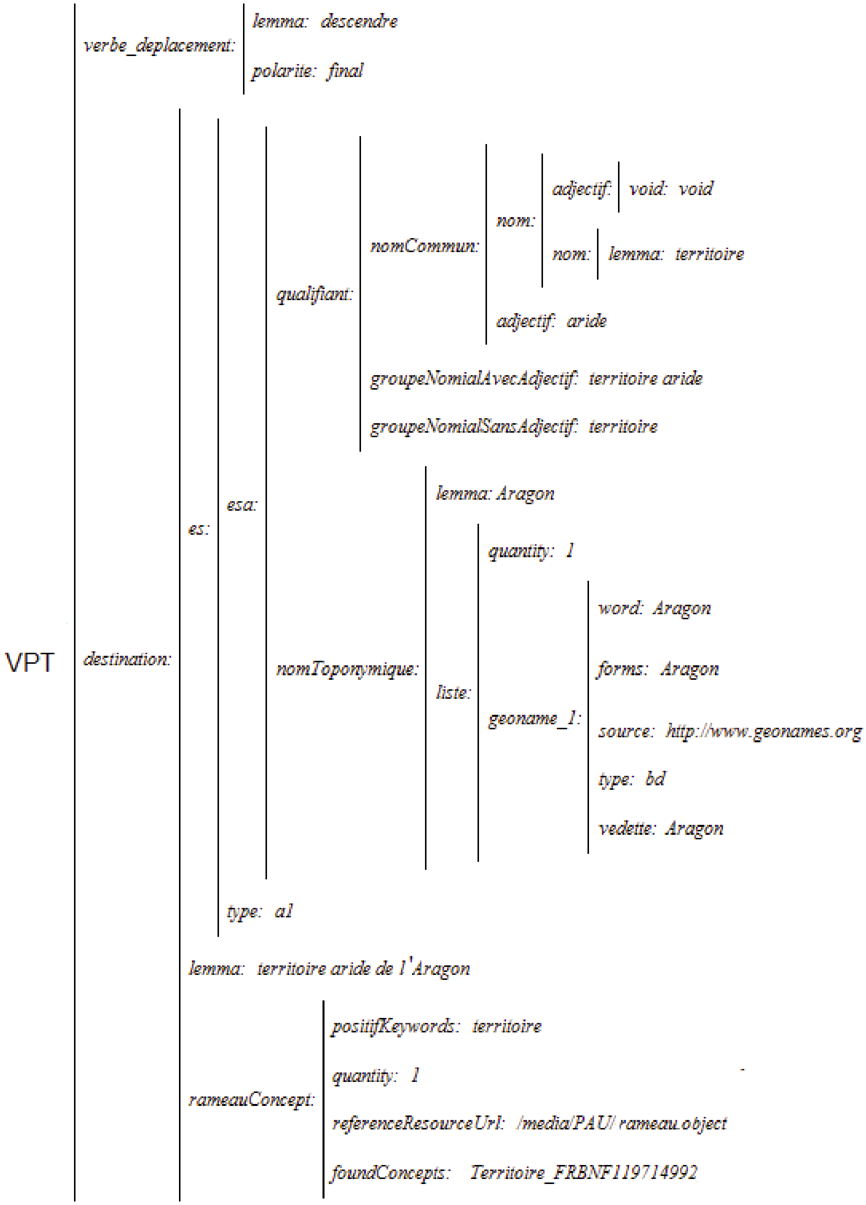
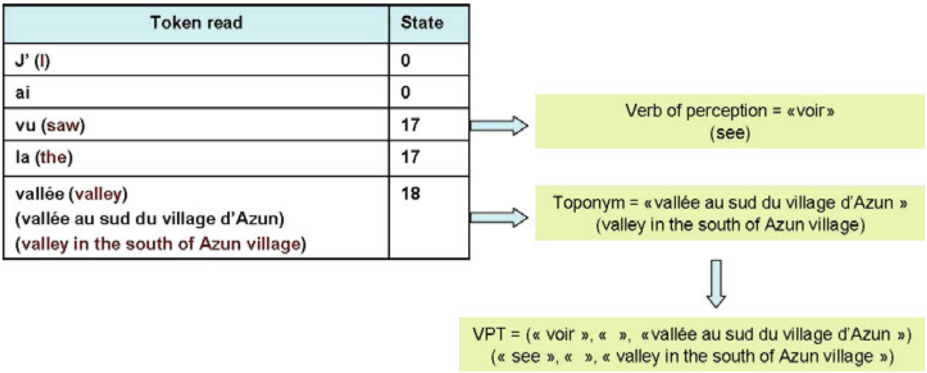


Fig. 3. Visualization of the triplet VPT marked

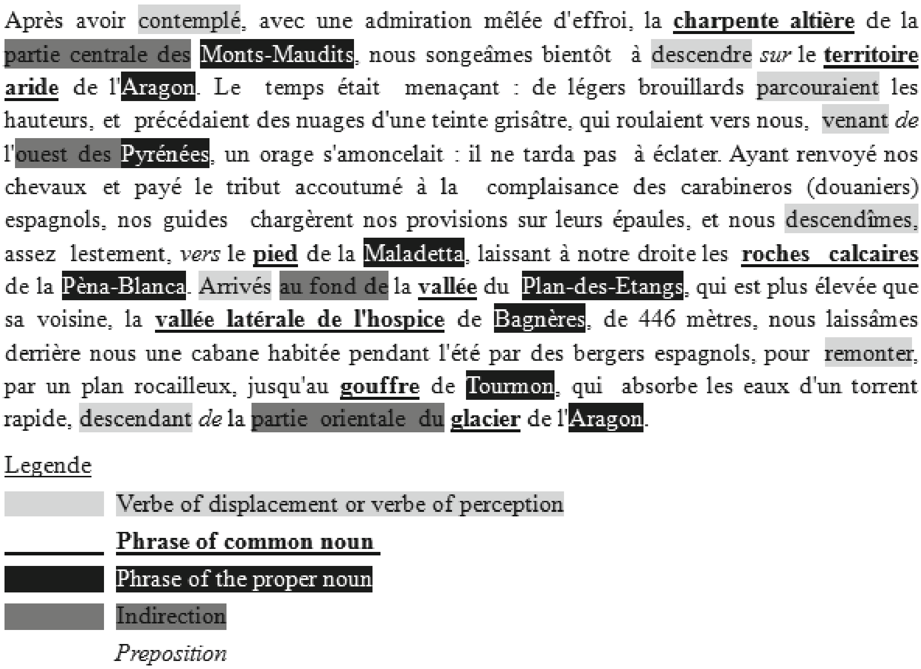
taking into account the specificity of this new language. For example, for English text, after the syntactic analysis also realized with TreeTagger. Concerning the marking process of verbs of movement and of perception, it lies on a specific lexical resource. And this lexical base is very specific for each natural language. According ([Tal00]), for the Romance languages like French or Spanish,



Token read	State
J' (I)	0
ai	0
vu (saw)	17
la (the)	17
vallée (valley) (vallée au sud du village d'Azun) (valley in the south of Azun village)	18

Verb of perception = «voir»
(see)
Toponym = «vallée au sud du village d'Azun »
(valley in the south of Azun village)
VPT = (« voir », « », «vallée au sud du village d'Azun »)
(« see », « », « valley in the south of Azun village »)

Fig. 4. The state transition of transducer for the second example sentence



Après avoir contemplé, avec une admiration mêlée d'effroi, la **charpente altière** de la partie centrale des Monts-Maudits, nous songeâmes bientôt à descendre *sur* le **territoire aride** de l'Aragon. Le temps était menaçant : de légers brouillards parcouraient les hauteurs, et précédaient des nuages d'une teinte grisâtre, qui roulaient vers nous, venant *de* l'ouest des Pyrénées, un orage s'amoncelait : il ne tarda pas à éclater. Ayant renvoyé nos chevaux et payé le tribut accoutumé à la complaisance des carabineros (douaniers) espagnols, nos guides chargèrent nos provisions sur leurs épaules, et nous descendîmes, assez lestement, *vers* le **pied** de la Maladetta, laissant à notre droite les **roches calcaires** de la Pèna-Blanca. Arrivés au fond de la **vallée** du Plan-des-Etangs, qui est plus élevée que sa voisine, la **vallée latérale de l'hospice** de Bagnères, de 446 mètres, nous laissâmes derrière nous une cabane habitée pendant l'été par des bergers espagnols, pour remonter, par un plan rocailleux, jusqu'au **gouffre** de Tourmon, qui absorbe les eaux d'un torrent rapide, descendant *de* la partie orientale du **glacier** de l'Aragon.

Legende

▒	Verbe of displacement or verbe of perception
___	**Phrase of common noun**
■	Phrase of the proper noun
▓	Indirection
	Preposition

Fig. 5. Automatic output colored text result of various lexico-syntactic processing

there are a large number of verb that indicates the direction of movement (eg "entrer", "sortir", "monter", etc..). In Contrary, for the Germanic languages like English or German, the direction is indicated by a particle (expressed by preposition) associated with the verb (e.g "go in ", "go out" "go up ", "go down"). This characteristic plays an important role to determine the construction of the transducer marking the triplet VPT. Finally the DCG rules to mark common noun, proper noun and toponym, must be rewrote. For example the case 3 of

Table 6. Some example with the term having instances in Rameau

Text	**Term**	**Instance in Rameau**	**Fathers**
Ayant atteint **le col ou le port** d'Albe, nous aperçûmes au-dessous de nous un petit lac de forme ovaloïde	port	Ports, Villes portuaires, Ports maritimes, Installations portuaires, Génie portuaire, Équipements portuaires...	Transports_maritimes, Terminaux_(transport, Ouvrages_hydrauliques, Canaux_(génie_hydraulique)
Nous prîmes le chemin du port de la Picade, en passant devant le **trou du Toro**	trou	Cavités, Orifices, Ouvertures (trous), Perforations	Surfaces_(mathématiques)
Nous ne regagnâmes nos **logements respectifs à Bagnères-de-Luchon** qu'après avoir été trempés jusqu'aux os	logement	Logements, Logement en milieu urbain, Hébergement, Habitat humain, Habitat (logement), Conditions d'habitation, etc.	Urbanisme
Nous songeâmes bientôt à descendre sur le **territoire aride de l'Aragon**	territoire	Territoires, Acquisition de territoire	Territoire_national

Table 7. Example marking the phrase of common noun for english text

```
%case 3 :(recursively)
group(nom1:N1..nom2:N2) --> commonNoun(N1), %hotel
        ls_token('of'), %of
        group(N2). %ville
commonNoun(adjectif:' '..nom:lemma:X) --> ls_token(_,
  lemma:X..stag:com, token).
...
```

the phrase of common noun(from line 6 to line 9 in the table 3) can be rewrote for English text as in the table 7.

4 Some Experimentations

Some global statistics We tried out our data processing sequence on a corpus of 14 books, in a nutshell we have:

- for 10555 occurrences of motion verbs found 1390 are involved in a VPT pattern.
- 560 VPT patterns containing candidates for sub-typing Place Name.

- 44 of them already exist in the domain-specific ontology,
- and 49 of them have matched with a key-concept in the RAMEAU thesaurus.

Verbs of perception effects. Thanks to the verbs of perception, we collect sentences such as those given in the example in table 8. It reveals new geographical information, which we could not take into account with the verbs of movement: *lac de Fachon, tour carrée de Vidalos, lac de Suyen.* We also have false positive response as with expressions like, *voir* ***la duchesse d'Albe***.

Table 8. A paragraph from the corpus "Travel stories" illustrating the use of verbs of perception

> Par 2.300 mètres, Wallon appuie à droite pour **admirer l'encadrement étrange et chaotique du petit lac de Fachon**. Nous sommes sur le meilleur observatoire pour **contempler l'énorme architecture du cirque de Troumouse**. **Contemplez la merveilleuse transparence des eaux du petit lac de Suyen ! s'écrie Russell**. Je suis allé **voir la duchesse d'Albe**. Le donjon de Lourdes **voyait les trois tourelles du château de Pau** qui **apercevait la tour carrée de Vidalos**.

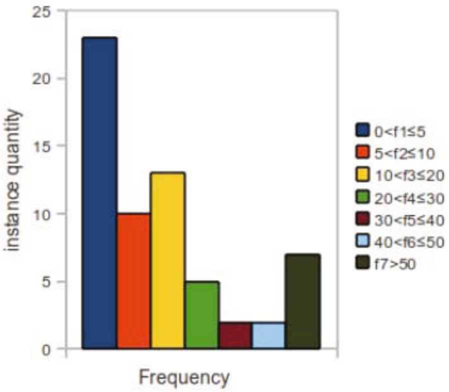


Fig. 6. Frequency of verbs operating in the linguistic structure: Verb of movement (Vmov) or Verb of perception (Vperc), preposition and toponym

We have counted 62 different occurrences of verbs. Among these verbs, the verb *voir* (to see) is the most used in our corpus. We created 7 classes of occurrences of the verbs (Figure 6) in relation to substantives. We notice that 7 verbs are in the frequencies of relations section **f7** *higher than 50* and that on the other hand, 23 verbs are in the frequencies of relations section **f1** *lower than 5 times.*

The occurrences of verbs are distributed according to Table 9. Among these verbs, 16 are verbs of perception.

Table 9. Distribution of verbs in our corpus

f1	**f2**	**f3**	**f4**	**f5**	**f6**	**f7**
abandonner	admirer	contourner	apercevoir	partir	atteindre	aller
approcher	contempler	diriger	entrer	passer	traverser	arriver
appuyer	dépasser	engager	revenir			conduire
border	entendre	franchir	venir			descendre
charmer	fixer	gagner	visiter			monter
dévorer	parvenir	observer				suivre
écouter	pénétrer	parcourir				voir
éloigner	redescendre	quiter				
envahir	regarder	rejoindre				
examiner	rentrer	remonter				
goûter		rendre				
grimper		retrouver				
longer		sortir				
marcher						
précipiter						
promener						
regagner						
réjouir						
repasser						
retourner						
rôder						
sentir						
toucher						

We finally get 214 distinct terms that are connected to verbs of movement, and 68 connected to verbs of perception. On these collections, 30% of terms appear only with verbs of perception, thus enabling us to widen the list of the potential candidates to the enrichment of ontology.

5 Conclusion

We have presented a global method for reducing ambiguity in complex geographic named entities. This method improves the task of geographic named entity annotation, both on identification and on sub-categorization. Thanks to a particular linguistic relationship, our main objective is to reduce the different opportunities that we can handle in the task of querying in huge external resources like generic thesaurus.

The assumption that we presumed on the presence of verbs of movement | verbs of perception as indicating a geographical connotation of the nominal group candidates linked to the place names is checked. The methodology suggested enables us to extract from our corpus of travel stories a lexicon of semantic labels.

The literature is quite poor in methods which focus on a deep determination of the sub-typing of place name. In [MTV07] when the named entities are classified and disambiguated, place name is assigned to the type "Location" or "Organisation". These place names have no details of their nature. In [RBH10], an ontology is used to reduced the ambiguity. However, This core ontology only defines a simple tree structure with four levels: a root (i.e., Earth), countries, states, and localities. Moreover, the pattern used to identify the sub-type candidate is very simple : for the cities in U.S, the pattern `[city-name, state-name]` is used; for all others `[name, country-name]` is used.

One of the advantages of our method is to use the resources with a large number of hierarchical concepts to reduce the ambiguity of sub-type for place name : the domain-specific ontology consists of more than 700 topographic concepts; the generic thesaurus RAMEAU is composed of more than 170 000 concepts in various domain. This allow to reduce the ambiguity of place name at various semantic level. Moreover, we use a generic pattern to identify the sub-type candidate of the place name : it's the pattern toponym (`[sub-type-candidate, indirection, place-name]`). Recursively, this pattern allows not only to extract the sub-type associated directly to the place-name (i.e., city in *Pau city*), but also determined the sub-type associated indirectly to it at different levels (i.e., hill in *argillaceous hill in the south of Pau city*).

The first objective of our work is to exploit key terms in several options:

– Either a term specifying the type during the geometry of the location recovery (for instance in resources like geographical databases or gazetteers) because a correspondence was found via geographical ontology.
– Or a term constitutes a proposal to the domain-specific ontology enrichment if it can't be found.
– Further option will be to use the pattern VPT to make validation *a posteriori*. Indeed, in a toponym we can have the place name not validated and whereas the sub-type is known: this can lead to a strong presumption that the place name is a real geographical named entity.

Acknowledgements. This work is realized in the framework of GeOnto project "Constitution, alignement et exploitation d'ontologies géographiques" (http://geonto.lri.fr/), partly funded by the French Research Agency (ANR-O7-MDCO-005).

References

[AM10] Abadie, N., Mustiere, S.: Constitution et exploitation dune taxonomie géographique á partir des spécifications de bases de données. Revue Internationale de Géomatique 20(2), 145–174 (2010)

[AS95] Asher, N., Sablayrolles, P.: A typology and discourse semantics for motion verbs and spatial pps in french. Journal of Semantics 2(12), 163–209 (1995)

[BLPD10] Brisaboa, N.R., Luaces, M.R., Places, A.S., Diego, S.: Exploiting geographic references of documents in a geographical information retrieval system using an ontology-based inde. Special Issue: Semantic and Conceptual Issues in Geographic Information Systems, GeoInformatica 14(3), 307–331 (2010)

[BOO87] Boons, J.-P.: La notion sémantique de déplacement dans une classification syntaxique des verbes locatifs. LANGUE FRANÇAISE, 5–40 (1987)

[DM00] Daille, B., Morin, E.: Reconnaissance automatique des noms propres de la langue écrite: Les récentes réalisations. Traitement Automatique des Langues 41(3), 601–621 (2000)

[FM03] Fourour, N., Morin, E.: Apport du web dans la reconnaissance des entités nommées. Revue Québécoise de Linguistique 32(1), 41–60 (2003)

[HEA92] Hearst, M.: Automatic acquisition of hyponyms from large text corpora. In: The Fourteenth International Conference on Computational Linguistics, Nantes, France (1992)

[JON94] Jonasson, K.: Le nom propre, constructions et interprétations, Duculot, Champs linguistiques (1994)

[LAU91] Laur, D.: Sémantique du déplacement et de la localisation en français: une étude des verbes, des prépositions et de leur relation dans la phrase simple. PhD thesis, Université de Toulouse II (1991)

[LEI04] Leidner, J.L.: Toponym resolution in text: "which sheffield is it?". In: The 27th, Annual International ACM SIGIR Conference (SIGIR 2004), Sheffield, UK, pp. 602–606. ACM Press, New York (2004)

[LES07] Lesbeguerie, J.: Plate-forme pour l'indexation spatial multi-niveaux d'un corpus territorialisé. PhD thesis, Université de Pau et des Pays de l'Adour (2007)

[LL07] Lee, S., Lee, G.G.: Exploring phrasal context and error correction heuristics in bootstrapping for geographic named entity annotation. Information Systems 32(4), 306–4379 (2007) ISSN 0306-4379

[Lou08] Loustau, P.: Interprétation automatique d'itinéraires dans des recits de voyage. type, Université de Pau et des Pays de l'Adour, address, month, note (2008)

[M.08] Aurnague, M.: Qu'est-ce qu'un verbe de déplacement?: Critéres spatiaux pour une classification des verbes de déplacement intransitifs du français. In: Congres Mondial De Linguistique FranÇAise, Paris, France (2008), doi:10.1051/CMLF08041

[MFP09] Maynard, D., Funk, A., Peters, W.: Using lexico-syntactic ontology design patterns for ontology creation and population. In: WOP 2009 Collocated with ISWC 2009 (2009)

[MTV07] Martineau, C., Tolone, E., Voyatzi, S.: Les entités nommées: usage et degrée de precision et de désambiguïsation. In: The 26th Conference on Lexis and Grammar, France (2007)

[MZB04] Malaise, V., Zweigenbaum, P., Bachimont, B.: Detecting semantic relations between terms in definitions. Ananadiou and Zweigenbaum, 55–62 (2004)

[Nan98] Chinchor, N.: Named entity task definition (version 3.5). In: The 7th Message Understanding Conference (MUC-7), Fairfax, VA (1998)

[RBH10] Roberts, K., Bejan, C.A., Harabagiu, S.: Toponym disambiguation using events. In: The 23rd Florida Artificial Intelligence Research Society International Conference (FLAIRS 2010), Applied Natural Language Processing track, Daytona Beach, FL, USA (2010)

[Tal00] Talmy, L.: How language structures space. In: Toward a Cognitive Semantics. The MIT Press, Cambridge (2000)

[TT97] Tversky, B., Taylor, H.A.: Langage et perspective spatiale. In: Sciences Cognitives, Masson, ch. 2 (1997)

[VJW07] Volz, R., Kleb, J., Mueller, W.: Towards ontology-based disambiguation of geographical identifiers. In: WWW 2007 Workshop I3: Identity, Identifiers, Identification, Entity-Centric Approaches to Information and Knowledge Management on the Web, Banff, Canada, May 8-12, pp. 1–7 (2007)

Author Index