THE GEOSPATIAL SEMANTIC WEB
Are GIS Catalogs prepared for This?

Carla Geovana N. Macário [1,2] and Claudia Bauzer Medeiros[1]
[1] *Institute of Computing, University of Campinas - UNICAMP, P.O.Box 6176, Campinas, SP, Brazil*
[2] *Embrapa Agriculture Informatics, Embrapa, P.O.Box 6041, Campinas, SP, Brazil*

Keywords: Geospatial Data, Geospatial Semantic Web, GIS Catalogs, Interoperability on the Web.

Abstract: Geospatial information catalogs are complex infrastructures that store and publish geographic information. They are an important part of Geographic Information Systems (GIS), systems that manage geospatial data for a wide variety of application domains. To be useful, a catalog must efficiently support discovery and retrieval of geospatial information, working as a key component for planning and decision-making in a variety of domains. Catalogs use standards to support data interoperability. However, the simple adoption of standards and specifications for geospatial data description enables only syntactic interoperability. Semantic heterogeneity still presents challenges for the so-called Geospatial Semantic Web. This work discusses some features that GIS catalogs should have, focusing in semantic issues. We tested some existing and well known catalogs, comparing them by means of these features. Based on this comparison, we identified some open issues that should be addressed considering advanced Geospatial applications on the Web.

1 INTRODUCTION

The term *geospatial data* refers to all kinds of data on objects and phenomena in the world that are associated with spatial characteristics and that reference some location on the Earth's surface. Examples include information on climate, roads, or soil, but also maps or telecommunication networks. Such data are a basis for decision making in a wide range of domains, ranging from studies on global warming to those on urban planning or consumer services.

For example, geographic applications for consumer services, like those provided by (Borges et al., 2007) and (Jones et al., 2003), assign a location to Web pages, based on existing geospatial evidences, such as addresses and phone number. This information can be subsequently used, for example, to find consumer services using fuzzy queries and to correlate Web pages spatially. In emergency management, geospatial information can be useful to identify areas prone to disasters (Klien et al., 2004) or to help in traffic control. In agriculture they are very useful for agroenvironmental planning (Macário et al., 2007; Macário and Medeiros, 2008), providing means to enhance agricultural productivity.

The Web plays an important role in this scenario, having become a huge repository of distributed geospatial information. Data are collected and stored by different organizations, which are required to exchange such data. These distributed data may be retrieved and combined in an *ad hoc* way, from any source available in the world, extrapolating their local context. Usually, the search for these data and methods is done by their syntactic content, focusing primarily in keyword matching. This can lead to retrieval of irrelevant data, and to omission of relevant facts. Hence, semantic interoperability is also a key issue in discovery, access and effective search for data in different application contexts. Solutions must take into account the constant modifications in the real world, and the evolution of our knowledge about the world.

There is a large amount of research on the management of geospatial data, including proposals of models, data structures, exchange standards and querying mechanisms. One area of activity concerns the so-called Geographic Information System (GIS) catalogs. These work as metadata catalogs that can be indexed by various means, such as by geographic location, and provide support for users to search for the data in different GIS data repositories. Catalogs are based on a common set of ideas which do not take semantic interoperability into account. This is a critical function necessary for advanced GIS applications, specially in the context of the Geospatial Semantic

MacÃ¡rio C. and Medeiros C.
THE GEOSPATIAL SEMANTIC WEB - Are GIS Catalogs prepared for This?.
DOI: 10.5220/0001839403350340
In *Proceedings of the Fifth International Conference on Web Information Systems and Technologies (WEBIST 2009)*, page
ISBN: 978-989-8111-81-4

Web (Egenhofer, 2002). In this work we identify important criteria that must be met by catalogs. Based on the results of comparing six widely used catalogs, we point out issues for research and development in the Semantic Web context. This discussion points at directions that must be followed in order to enhance the interoperability of GIS on the Web.

2 RELATED CONCEPTS

2.1 Geospatial Semantic Web

The Semantic Web was initially proposed by Berners-Lee (Berners-Lee et al., 2001) as a way to bring structure to the meaningful content of Web pages, creating an environment where users can obtain information based on semantics and not only in syntax. In this scenario, the Semantic Web would enable machines to comprehend semantic documents and data, through: (1) adoption of standardized data element names to describe and exchange the data; (2) description of information in terms that allow common understanding; (3) exposing data to be found and retrieved; (4) designing efficient retrieval mechanisms.

A standard establishes the name of data elements (metadata) and/or groups of these elements, providing a common set of terminology and definitions for the description and exchange of data. The adoption of a common vocabulary in this description ensures that data producer and consumer share the same understanding of data. Hence, in the Semantic Web, the description of the meaning of data using ontology terms, through standardized metadata is a way to provide semantics, increasing interoperability. This description process is called *annotation*.

The Semantic Web for geographic information, called Geospatial Semantic Web by Egenhofer (Egenhofer, 2002), is a way to process requests involving different kinds of geospatial information. This requires the development of multiple spatial and domain ontologies, their representation in a way that computers can understand and process, the processing of queries considering these ontologies and the evaluation of results based on the required semantics. All of this leads to the search for a geospatial information retrieval framework that relies on ontologies, allowing users to retrieve desired data based on their semantics.

In spite of several efforts, the Semantic Web is far from becoming a reality (Shadbolt et al., 2006). Although several standards have been developed and adopted, there are too many variables that need to be considered. The variety of user profiles and needs, and of application domains – and thus of ontologies – are just some of these factors. So far, most retrieval engines are restricted to text, and other kinds of media pose countless challenges to the effective implantation of the Semantic Web (Macário and Medeiros, 2008).

2.2 Geospatial Catalogs

Catalogs are complex structures that enable data to be found and retrieved, through the publishing of descriptions of these data by metadata, known as annotations (Nogueras-Iso et al., 2005), and operations on these annotations. Catalogs offer search mechanisms that access them to retrieve the desired data.

A GIS catalog is a Web application to publish descriptions of geospatial data, enabling users to search for the desired data (OCG, 2006). Because of standardized interface specifications, different users can access them from all kinds of sites to search for the content they need.

The Open Geospatial Consortium, OGC (OCG, 2006) is a non-profit international organization that is leading the development of standards for geospatial and location based services. The consortium aims at interoperability among geospatial systems, making complex spatial information and services accessible and useful to all kinds of applications. It describes three basic operations that a geographic catalog should provide: publication, discovery and retrieval of geospatial metadata.

Geospatial data is described by metadata and these descriptions are published in a catalog to support data discovery. Data discovery can be performed either by browsing the content of the catalog or by choosing certain query terms. Once the desired metadata is found, the data referenced can be retrieved.

3 DESIRABLE GIS CATALOG FEATURES

In a Web environment, GIS users need to explore available databases to discover the desired information. In order to find the data, the first step is to search for specific GIS catalogs and, once connected to the catalog, look for candidate metadata describing the desirable data. As the needed data is found, the users can download and use it in theirs applications.

However, this is not an easy task to perform. Geospatial data are complex, due to their spatial component and its dynamic characteristics. Besides this, users are hampered in their queries because of the many different concepts and terms used to describe data items. Catalogs seldom publish semantic annotations. One possible approach for this is the use of terms of an ontology to describe data, helping to remove the ambiguity. The increase in quality of the retrieved information and enhanced interoperability are some benefits from the adoption of semantic descriptions, also known as semantic annotations. Although there is extensive research in geospatial semantics, it is focused mainly in the adoption of standardized data element names and of ontology terms to describe the data. It is not common to find semantic catalogs, which are those that publish semantic annotations and support search on them as a way to enhance the retrieval of information. In this section we describe the main features that a catalog should provide in order to make the Geospatial Semantic Web a reality. These features are based on those presented by (Larson et al., 2006) and (ESRI, 2003), always considering the user viewpoint.

Feature 1: OGC Compliance. One of the many standards proposed by OGC is the Catalog Services Interface Standard (CAT), which supports the ability to efficiently publish and search collections of metadata about geospatial data, services and related resources. Hence, focusing in interoperability, a catalog should be OGC compliant, enabling its use by users and also by other catalogs.

Feature 2: Standards for Metadata. Catalogs should support metadata standards. The growing need for geospatial information led to the development of a number of initiatives to obtain spatial metadata according to a variety of formats within agencies, communities of practice, or groups of countries. This resulted in well established and widely used standards like the ISO 19115 Metadata Standard (ISO, 2008), or the FGDC geospatial metadata standard (FGDC, 1998). The objective of these standards is to provide a common set of terms and definitions for the documentation and exchange of geospatial data.

The ISO 19115 standard (ISO, 2008) is a well known standard for geographic information metadata that defines the schema required for describing geographic information and services. It provides information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data (Silva, 2008). The Federal Geographic Data Committee (FGDC, 1998) develops geospatial data standards for implementing the USA National Spatial Data Infrastructure. The Content Standard for Digital Geospatial Metadata (CSDGM), which is often referred to as the FGDC Metadata Standard, provides the definition of profiles and extensibility through user defined metadata extensions.

Feature 3: Support Advanced Search. Catalogs should provide different means for users to perform their queries, considering different access levels to each catalog and its contents. Users may perform the search considering specific metadata elements, in a way to refine their query. It is a good choice to provide exploration tools, enabling users to explore the retrieved data to determine suitability to their applications. Users should be able to select the desired sources and categories and kinds of data to be retrieved. Besides this, it is important that each search option be described, enabling its use by foreign people. In this sense, the adoption of standard interfaces can be very useful. Catalogs should also allow users to view metadata records to determine if the retrieved data is suitable for the intended use.

Feature 4: Save Data Online. Catalogs should allow users to view entire metadata records to determine if the corresponding data is suitable for the intended use. Once the user finds the desired content in a catalog, it is important to have means to save its description or even the content itself. Hence, catalogs should support a range of methods for online data delivery (e.g., live data streaming, commonly used data formats, FTP download, and CDROM).

Feature 5: Provide Access to Multiple Servers. A catalog should support search considering other metadata servers, increasing the number of repositories to be searched. It has to be done in a consistent way, enabling users to discovery new information repositories. The study (ESRI, 2003), shows that most users do not perform distributed search due to problems on catalogs. Instead, they go to specific GIS catalogs and browse them to find relevant data for their projects. The portal should also support a search against a single catalog.

Feature 6: Cater to Geospatial Data Diversity. Geospatial data users are always looking for different kinds of data, and also Web services. Hence, catalogs should provide description of all these kinds of data, allowing access to them. For example, maps should be viewable in the browser or through an appropriate software.

Feature 7: Support Semantic Search. Traditional search mechanisms based on keyword matching are restrictive. More expressive search algorithms, which enhance recall and precision, should be available – e.g., via thesauri, gazetteers and multilingual processing. A more flexible option is the use of ontology terms to describe the data. In this sense, the catalog should enable automatic matching of these terms during the discovery process.

4 COMPARING GIS CATALOGS

4.1 Overview of Selected Catalogs

We tested some GIS catalogs, as a means to identify issues for research and development in the Semantic Web context in order to enhance the interoperability of GIS on the Web. Although these catalogs are standardized interface specifications, they are implemented considering different requirements, even for the geographic domain. In this test we considered the guidelines we stated in section 3.

Embrapa Information Agency (Souza et al., 2006) is a Brazilian Web system to organize, deal with, store, publish and access the technological information generated by Brazilian Agricultural Research Corporation - Embrapa and other agricultural research institutes. Knowledge is organized hierarchically, under the form of a tree. Although directed to agricultural domain, knowledge is described using Dublin Core metadata (Weibel et al., 1995), to allow its retrieval by different user profiles. Only a syntactic search for discovery of the stored resources is available, and search results can be saved in a textual file.

INSPIRE (www.inspire-geoportal.eu) is an European initiative that aims to provide geospatial information to be used to formulate, perform and evaluate european policies. Its objective is to create a spatial information infrastructure to deliver integrated spatial information services. The main users of INSPIRE include policy-makers, planners and managers at European, national and local level as well as the citizens and their organisations.

FAO – The UNO Food and Agriculture Organization leads international efforts to defeat hunger (FAO, 2008). The FAO catalog aims to share geographically referenced thematic information between different organizations. It was implemented using the GeoNetwork opensource (geonetwork-opensource.org), a standards based, free and open source catalog application to manage spatially referenced resources through the web. It offers metadata editing and search functions, as well as an embedded interactive web map viewer. The catalog provides access to interactive maps, satellite imagery and related spatial databases maintained by FAO and its partners.

IDEE – Spatial Data Infrastructure of Spain (www.idee.es) aims to integrate all data, metadata, services and geographic information produced in Spain. Its goal is to make the location, identification, selection and access of these contents an easier operation to their potential users. The IDEE catalog enables users to search for geographic information – maps, ortophotos, etc – available for an area or a theme, in a specific period of time.

GeoSpatial One Stop - GOS (gos2.geodata.gov) is a public GIS catalog that aims to improve the access to geospatial information and data. The catalog is constructed under the U.S. Geospatial One-Stop E-Government initiative for enhancing government efficiency and improving citizen services. Through the catalog it is possible to find data or map services, make a map, browse community information, cooperate on data acquisition. Information is provided by government agencies, individuals, and companies, or obtained by harvesting the data from geospatial clearinghouses.

4.2 Comparison of Catalogs

Table 1 shows a comparative analysis of the presented catalog systems, taking into account the features presented on section 3.

Except for the Embrapa Agency and GOS, all the analyzed tools were implemented considering the specifications provided by OCG. Though GOS is not compliant with OGC, it was implemented according to the National Spatial Data Infrastructure provided by FGDC, which also focus on cooperative production and sharing of geographic data. All the catalogs provide data that are described using metadata standards, most of them using FGDC or ISO 19115. This indicates that they all aim to promote the exchange of the data they provide. However, to really support data exchange, it is necessary that these descriptions be supplied in an exchangeable format, like XML or csv. The translation of element names from a standard, or saving data descriptions in a textual format, as Embrapa Agency and IDEE do, restricts this exchanging.

The search for data is provided both in simple and in advanced ways in all tested catalogs, except on IDEE, which offers only the advanced one. A simple search enables the user to look for the keyword occurrence within the entire record. However, this can be a hard operation. Embrapa Agency, though offering both kinds of search, has a limited number of options for the advanced search. The same occurs with IDEE. Only three of the catalogs provide access to multiple GIS catalogs, supporting search in different repositories. Though IDEE has this feature, at present it accesses only the National Geographic Institute data. All catalogs provide digital and non digital data, but INSPIRE also provides search for services and applications, which can improve the interoperability among geographic systems. Finally, none of the systems analyzed enables a search based on the semantics of the data.

Table 1: Evaluated GIS Catalogs.

Catalog	OGC Compliance	Standard Metadata	Save data	Advanced Search	Multiple servers	Data Diversity	Semantic search
Embrapa Information Agency (Brazil)	no	Dublin Core, in portuguese	Yes, in a textual format (descriptive)	yes	no	Digital and non digital	no
INSPIRE (Europe)	yes	ISO19115	no	yes	yes	Digital, web services and applications	no
FAO Catalog (FAO)	yes	ISO19115 FGDC Dublin Core	Yes, the standard metadata in XML format	yes	yes	Digital and non digital	no
IDEE (Spain)	yes	ISO19115 in different languages	no	no	no	Digital and non digital	no
Geodata.gov (USA)	yes	FGDC ISO19115	yes, in textual format (csv)	yes	yes	Digital and non digital	no

5 OPEN RESEARCH TOPICS

This section summarizes some open research issues that we have identified as a result of the comparison presented on subsection 4.2. This reflects what we expect to be the most important features to be supported by catalogs, towards making the Geospatial Semantic Web a reality.

- *Search on Multiple Servers.* We identify this as a challenge because of the following: (1) some catalogs presented bad performance, thus motivating the need to develop or adopt better algorithms; (2) some results were very difficult to interpret because of the language they use, making the data useless. Hence, content description has to be also in a well-known language; (3) some results were dependent on available services. As many catalog or data providers were offline, it was impossible to get the data.
- *Semantic Search.* This is a central issue to be considered. The available catalogs do not provide this kind of search, in spite of its usefulness when it comes to geospatial data. A good survey of semantic search approaches can be seen in (Mangold, 2007).
- *Query Modification.* Although this is part of the previous item, it is also an important issue to be considered by itself. Query modification in catalog search can help disambiguate search expressions and enhance semantics.
- *Adoption of Standards.* This is a large ongoing effort, focusing on interoperability of geospatial data. The FAO Catalog and GOS are good examples for this issue. However, each one is based on a different, but well known, geospatial standard. Hence, if their contents are to be combined, one must develop translators from one to the other. Common standards would avoid this kind of problem.
- *Standard Interfaces.* Once a user wants to search for data in different catalogs, she has to identify the available search options and what each field means. We identify the design of common interfaces as a promising research area. The development of standardized services can also enhance the use of the available catalogs.

6 CONCLUSIONS

Geospatial data available on the Web are very useful to answer important questions for various domains, such as emergency management, services and agroenvironmental planning. Geographic catalogs are organized as descriptive lists of metadata, which describe existing geospatial data. Through the publishing of these metadata, users are allowed to search for the desired information to be used in their systems. However, this search is not a trivial task, subject to a wide range of problems. In particular, in the context of the Geospatial Semantic Web, there are two main issues to be addressed: (1) how to perform semantic search, seen as a means to reduce the ambiguity of terms? (2) what should be done in order to have a huge semantic geospatial data network?

This work discussed features that GIS catalogs should present, focusing in the Geospatial Semantic Web. These features are based on interoperability issues, from the user viewpoint. We tested some existing and well known GIS catalogs, comparing them by means of these criteria. Furthermore, we identified research and development issues that are not addressed by the tested catalogs, and that are very important for advanced Geospatial applications. Although many of the existing catalogs are good, they are far from what is needed to support Semantic Networks. Much effort has to be directed to the use of ontologies on search operations. Distributed search also represents a challenge, as this is not a controlled operation. Finally, the adoption of standard interfaces could facilitate the search for data. Initiatives such as OGC are doing a good work in this direction. However there are still gaps to be filled.

ACKNOWLEDGEMENTS

We would like to thank Embrapa, FAPESP, Virtual Institute FAPESP-Microsoft (eFarms project), CNPq and CAPES for the financial support for this work.

REFERENCES

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. *Scientific American*, pages 34–43.

Borges, K. A. V., Laender, A. H. F., Medeiros, C. B., and C. A. Davis, J. (2007). Discovering geographic locations in web pages using urban addresses. In *GIR '07: Proceedings of the 4th ACM workshop on Geographical information retrieval*, pages 31–36, USA. ACM.

Egenhofer, M. J. (2002). Toward the semantic geospatial web. In *Proc. of the ACM GIS02*, pages 1–4.

ESRI (2003). Implementing a metadata catalog portal in a GIS network. Technical report, ESRI.

FAO (2008). *FAO - GeoNetwork*. http://www.fao.org/geonetwork/srv/en/main.home.

FGDC (1998). *FGDC-STD-001-1998. Content Standard for Digital Geospatial Metadata*. Washington, D.C.

ISO (2008). *ISO 19115:2003 Geographic information – Metadata*. ISO. Available on:<http://www.iso.org/iso/home.htm>.

Jones, C. B., Abdelmoty, A. I., and Fu, G. (2003). Maintaining ontologies for geographical information retrieval on the web. In *OTM Confederated International Conferences - CoopIS, DOA, and OOBASE*, pages 934–951.

Klien, E., Einspanier, U., Lutz, M., and Hbner, S. (2004). An architecture for ontology-based discovery and retrieval of geographic information. In *7th AGILE Conference on Geographic Information Science*, Heraklion, Greece. Parallel Session 3.1– Semantics I.

Larson, J., Siliceo, M., Silva, M., Klien, E., and Schade, S. (2006). Are geospatial catalogues reaching their goals? In *9th AGILE Conference on Geographic Information Science - Poster*, Visegrd, Hungria.

Macário, C. G. N. and Medeiros, C. B. (2008). A framework for semantic annotation of geospatial data for agriculture. *Int. J. Metadata, Semantics and Ontology - Special Issue on "Agricultural Metadata and Semantics"*. Accepted for publication.

Macário, C. G. N., Medeiros, C. B., and Senra, R. D. A. (2007). The webmaps project: challenges and results (in portuguese). In *IX Brazilian Symposium on GeoInformatics - Geoinfo 2007*, Brazil.

Mangold, C. (2007). A survey and classification of semantic search approaches. *Int. J. Metadata, Semantics and Ontology*, 2:23–34.

Nogueras-Iso, J., Zarazaga-Soria, F., Bjar, R., lvarez, P. ., and Muro-Med, P. (2005). OGC catalog services: a key element for the development of spatial data infrastructure. *Computers & Geosciences*, 31:199–209.

OCG (2006). CSW 2.0 FGDC application profile. Technical Report OGC 06-129r1.

Shadbolt, N., Berners-Lee, T., and Hall, W. (2006). The semantic web revisited. *IEEE Intelligent Systems*, 21(3):96–101.

Silva, H. (2008). *MIG - Metadata for Geographic Information –Introduction to ISO 19115 standard*. Portuguese Geographic Institute, Portugal. (in portuguese).

Souza, M. I. F., Santos, A. D., Moura, M. F., and Alves, M. D. R. (2006). Embrapa information agency: an application for information organizing and knowledge management. In *II Digital Libraries Workshop*, pages 51–56, Brazil. (in portuguese).

Weibel, S., Godby, J., Miller, E., and Daniel, R. (1995). OCLC/NCSA Metadata Workshop Report. Web site http://www.oclc.org:5046/oclc/research/conferences/metadata/ dublin_core_report.html.