Are Geospatial Catalogues Reaching their Goals?

Jennifer Larson[1,2], Maria Antonia Olmos Siliceo[1,3], Marcelino Pereira dos Santos Silva[1,4,5], Eva Klien[1], Sven Schade[1]

[1]Institute for Geoinformatics –University of Münster, Germany
[2]Department of Geography - San Diego State University, USA
[3] Instituto Tecnológico de Toluca - 52140 Metepec, México
[4]National Institute for Space Research - São José dos Campos, Brazil
[5]Department of Informatics - Rio Grande do Norte State University, Brazil

jlarson@rohan.sdsu.edu, maria_olmos@yahoo.com, mpss@dpi.inpe.br, klien@uni-muenster.de, schades@uni-muenster.de

SUMMARY

The goal of a geospatial catalogue is to support a wide range of users in discovering relevant geographic information from heterogeneous and distributed repositories. Syntactic interoperability is secured by implementing the OGC web catalogue specification and compliant metadata standards. The prime challenges thus are to provide suitable access points and search functionalities on metadata, so that users with different experiences and different viewpoints are able to conduct efficient catalogue queries. In this paper we present a survey on existing geospatial web catalogues with the goal of evaluating their discovery functionalities. Based on the results of the survey we propose improvements, which we think are crucial to overcome the ongoing frustration and ineffectiveness of currently available catalogue search functionalities.

KEYWORDS: *geospatial metadata, OGC web catalogue, geographic information discovery*

INTRODUCTION

The increasing relevance of information sharing, notably with geodata in an open environment such as a Spatial Data Infrastructure (SDI), has posed new issues and new challenges in data exploration. Geospatial catalogues publish metadata on geographic information and provide mechanisms for querying and retrieving information from distributed repositories (Bernard et al., 2005). They are designed to provide access points to a wide range of users that demand specific geodata for a wide range of applications. However, catalogue users often have frustrating search experiences caused by unclear access points, ambiguous search methods, unsuitable metadata, and long response times.

The goal of this paper is to evaluate the discovery functionalities of Geospatial Catalogue Services, as well as making proposals for improvement. Considering the strategic aspects of information sharing and retrieval through catalogues, we concentrate our efforts on two subjects: (1) finding access points to geospatial metadata in order to perform searches, and (2) assessing the discovery potential of search methods offered to find information that meets the user needs. To evaluate the points above, a survey was performed to find out if geospatial catalogues are really reaching the desired goals. We introduce a scenario to illustrate some of the drawbacks found while searching for metadata in different catalogues. Based on the observations and results acquired during the survey, relevant topics on interface design, metadata quality, thematic search methods and related semantic issues to improve catalogue functionalities are discussed.

We first give an overview on the relevant standards, specifications and functionalities of geospatial catalogues. This is followed by a description of the survey settings and a subsequent discussion of the results. In the last section, proposals on how to improve discovery in catalogues are made.

GEOSPATIAL CATALOGUE FUNCTIONALITIES

The main functionality of catalogue services is to publish meta-information on geospatial resources and to provide methods to search and query that information (Bernard et al., 2005). The concept of geospatial catalogues is based on the sharing of information across distributed networks. The Open Geospatial Consortium (OGC, http://www.opengeospatial.org) works in conjunction with the International Standards Organization (ISO, http://www.iso.org) to develop standards to achieve interoperability for geographic information and services.

In order for a geospatial catalogue to be considered OGC compliant, it is obligated to follow the OGC Catalogue Services Specification (OGC, 2005). Geospatial metadata is designed for the communication and discovery of all types of information. Standards for geographic metadata are defined in ISO 19115 (Geographic information –Metadata) and ISO 19119 (Geographic information - Services) (ISO/TC-211, 2003; ISO/TC-211, 2005).

FRAMEWORK OF THE CATALOGUE SURVEY

The survey on OGC catalogs was completed in November 2005 with the purpose of evaluating the discovery functionalities of geospatial catalogues. Selected from a list of twenty, five OGC compliant geospatial catalogues were chosen for assessment:

- **INSPIRE Geo-Portal** (http://eu-geoportal.jrc.it/): INSPIRE (Infrastructure for Spatial Information in Europe) is an initiative launched by the European Commission and developed in collaboration with Member States and accession countries. The INSPIRE Geo-Portal is Europe's Internet access point for Spatial Data and Services.
- **IDEC Catalog** (http://www.geoportal-idec.net): IDEC (Infraestructura de Dades Espaciales de Catalunya) is the Catalan Spatial Data Infrastructure Project.
- **Gigateway** (http://www.gigateway.org.uk/): Gigateway is a free web service aimed at increasing awareness of and access to geospatial information in the UK.
- **GSDI Gateway** (http://gsdi.org/Default.asp): The GSDI (Global Spatial Data Infrastructure) Association promotes international cooperation and collaboration in support of local, national and international SDI developments.
- **GEODATA. GOV** (http://www.geodata.gov): geodata.gov is a geographic information system (GIS) portal that serves as a public gateway for improving access to geospatial information and data under the US Geospatial One-Stop e-government initiative.

The catalogues were evaluated based on four key criteria:

1. *Direct access* to the catalogue: relates to the user interface of the website and whether it is designed to allow for direct access to the catalogue. Is the catalog clearly accessible from its homepage and simple to navigate? The number of clicks needed to get to the actual catalogue client interface is one of the indicators to measure its accessibility.
2. *Suitable search and discovery methods*: Is the user allowed to define the query based on spatial, temporal, attribute, or categorical parameters? This could be carried out by various menu options and/or a mapping service.
3. *Metadata retrieval*: concerned with the accessibility of the metadata from the list of search results. Which items are displayed to the user? To meet the criterion for metadata retrieval the

catalogue must provide a summary of the relevant metadata and or a working link to the complete metadata. The actual geodata should be accessible by a link contained in this metadata.

4. *Retrieval time*: the metadata had to be accessible in a speedy manner; the degree is indicated with *poor* (i.e. "user could not bother to wait any longer"), *regular* ("reasonably fast responds"), and *good* ("surprisingly fast putting a smile on the users face").

To illustrate the functionalities that were focused on and the findings of the survey, consider the following scenario that we have conducted with all five catalogues. The purpose of the survey was to assess the search and discovery functionality of the catalogues, not the degree to which the results are suitable. Suitability is defined by individual users and therefore is not a primary aspect for evaluation. The problems that our prototypical user Robert is experiencing in this example are a synthesis of problems discovered in the different catalogues during the survey. The specific results for each catalogue will be presented and discussed in the next section.

Robert has just been hired by the City as a consultant to investigate the affects of the conversion of rural to urban land uses on water quality over the last five years. He was able to obtain the necessary land use data and is now collecting water quality data. Robert knows that there has been water quality data collected in the area for many years and that it should be available online. Robert locates a gateway for geospatial information that provides a link to an online geospatial catalogue. The link takes him to a page that he finds a little confusing, but after changing the language preference, he is able to locate and open the web catalogue service interface. A mapping service allows him to define the spatial extent of the search; but it fails to load. Roberts decides to do without the spatial extent and continues to define his query by entering the keyword *water quality* into the query interface. The first search immediately returns no results and Robert discovers that he has misspelled the keyword. He makes the correction and restarts the search. This time he additionally selects a predefined topic category from a drop down menu to narrow the search, even though he is not completely confident if the selected category of *Inland Waters* actually covers the kind of information he is looking for. This search takes almost a minute before the results are listed, but Robert discovers that the long list of registries only provides him with a couple of results that could be useful to him. To avoid having to scan through the whole list for suitable results he redefines the search using additional keywords and restricts the time period to the last five years. This process is iterated a few times with different terminology and different combinations of parameter, until a set of seemingly useful registries is returned. When accessing the full metadata, Robert discovers that only one provides online access to suitable geodata. The other registries turn out to be outdated, incomplete, containing dead links or covering the wrong spatial extent.

WHY IS DIFFICULT TO RETRIEVE INFORMATION FROM CATALOGUES?

Geospatial catalogues are designed to provide access points to users who demand specific geodata for a wide range of applications. However, the survey has shown that the use of geospatial catalogues causes problems to arise: unclear access points, ambiguous search methods, unsuitable metadata, and long response time are some of the difficulties found. When users try to locate data that might be appropriate to meet their demands, these problems reduce the catalogue usage potential. The results of the survey are summarized in Table 1 and discussed in the remainder of this section.

When dealing with the task of finding appropriate data for their demands, users need a direct and objective interface to allow them to go straight to the point. When trying to get access to the actual catalogue client some interfaces turned out to be inappropriate. The number of clicks required to reach the search interface show how deep the navigation hierarchies are. Also, the variety of link labels that take you to the search interface is remarkable: *Data Catalog*, *Data Locator*, *Search for Data*, *Electronic Gateways*, and others. Besides the unclear labeling, excess of information and uncorrelated links made the navigation difficult. Among the evaluated web catalogues, the INSPIRE

Geo-Catalog (*Figure 1*) offered the most structured interface with a clear focus on its actual functionality (i.e. search for geospatial resources). In contrast, the user interface of GEODATA.GOV displays a huge amount of only indirectly related information, which makes it difficult to concentrate on the actual task. Errors were experienced while loading the catalogue user interfaces due to missing plug-ins and a lack of proper information on where to get them was noted (e.g. mapping service of GSDI Gateway).

Table 1: Catalogue Survey Results (X indicates that the criteria has been met)

Catalogue	Direct Access	Search Methods				Metadata Retrieval	Retrieval Time
		Free Text	Category	Temporal Period	Spatial Extend		
INSPIRE - Geo-Portal	1	X	X	X	X	X	☺
IDEC Catalog	1	X	X		X	X	😐
Gigateway	1	X	X	X	X	X	☺
GSDI Gateway	3	X	X	X	X		☹
GEODATA.GOV	0	X	X	X	X	X	☺

The search and discovery methods include free text, controlled vocabulary (category) search, location/spatial extent, time period, theme, provider and other parameters input to defining the search. A sufficient set of search parameters were offered by all of the examined catalogues in the survey, but major difficulties were experienced when making use of them. The variety of criteria provided by search interfaces created confusion. For example, predefined categories are offered for search but they are not ordered (e.g. INSPIRE Geo-Portal, Figure 1) and the specific controlled vocabularies of the catalogues introduce different terms for the same category (e.g. *hydrology*, *inland waters*, *inland water resources*, *water bodies*). This is complicated by the fact that none of the catalogues offers further explanations on these categories. As described in the example scenario, the thematic free text search on heterogeneous resources leads to further problems: e.g. keywords chosen by the user might not fit the metadata entries, category, theme or other criteria does not fit exactly the features available. Consequently, none or just a few registries are found (low recall), or hundreds of unsuitable items are returned (low precision). In both cases, ambiguity is usually present and the user is challenged to keep trying new sets of criteria, or has to waste time examining the multitude of registries returned.

More obstacles were faced when retrieving metadata from the list of results. For example, outdated or insufficient information and irregular descriptions complicate the ability to evaluate, if the described resource could fit the user's purpose. Metadata that is not compliant with current standards and issues with multiple languages are further obstacles that interfere with quick and precise access to geodata. Only the IDEC Catalog provides a search interface in multiple languages with almost all

metadata being described in English, Catalan and Spanish. An integrated thesaurus on a list of topics even provides translation between these languages.

***Figure 1*:** User interface for the catalogue search from the INSPIRE Geo-Portal.

Finally, the issue of performance should not be underestimated. Even if search functionalities are brilliant, this fact will remain unknown if the respond time takes too long. This is a subjective criterion, but as shown in Table 1, the retrieval procedure for metadata containing data access information sometimes took longer than one normally would be willing to wait.

PROPOSALS TO IMPROVE USABILITY

Are geospatial catalogues reaching their goals? Considering that the strategic role of geospatial catalogues is to provide suitable access points and search functionalities, so that a wide range of users are able to conduct efficient catalogue queries, we observed that current implementations fail on some crucial requirements. From our point of view, the most relevant ones are related to the interface design, metadata quality, thematic search methods and related semantic issues. In this section we discuss current challenges including proposals for improvement.

The survey underlines the need for direct access to query interfaces and illustrates that some of the currently available catalogues still lack this ability. The relevant topics must be presented in a straightforward manner, without confusing links, additional text or uncorrelated information. Key topics on interface design approached by software engineering (Shneiderman & Plaisant, 2004) must be taken into account in catalogue implementations. Specifics of these topics within the web

environment (Nielsen, 1999) and usability criteria particularly related to the geo-web (Wachowicz et al., 2005) require special attention. Cantán et al. (2003) specifically assess the usability of query formulation in geospatial catalogs and propose to allow query formulation on a more abstract level. Depending on the cross-national scope, appropriate language features on the interface level should be available either by actively including the user (offering multiple choices) or by inferring a language that the user understands.

Successful catalog search depends on the quality of the metadata records. The issue of metadata appropriateness can be derived from factors related to its creation and maintenance. Metadata is generally created by humans, which brings uncertainty and incorrectness. To reduce drawbacks related to metadata generation automatic methods are needed as proposed in Manso et al. (2004). Another suggestion concerns regular metadata update, in order to avoid some of the technical errors discovered in the catalogue survey. Feedback mechanisms, which allow users to inform metadata administrators about problems, would help to eliminate errors, inconsistencies and quality problems (Craglia & Evmorfopoulou, 2000).

The effectiveness of a catalogue query highly depends on the methods implemented for discovering relevant information in the registered metadata. Even though the use of thesauri and natural language processing techniques (e.g. Richardson & Smeaton, 1995) might increase the semantic relevance of search results, keyword-based search is inherently restricted by the ambiguities of natural language. As a result, keyword-based search can have low recall if different terminology is used and/or low precision if terms are homonymous or because of their limited possibilities to express complex queries (Bernstein & Klein, 2002). Thus specific semantic techniques for knowledge description and discovery are demanded for geospatial catalogues. One proposal to avoid problems of metadata inappropriateness and semantic heterogeneities within search mechanisms is the Semantic Geospatial Web (Egenhofer, 2002), which approaches real machine-readable descriptions, allowing users to retrieve more precise data and to count on information with semantic issues attached. Ontologies have been identified as a valuable means to formally and explicitly capture the meaning of terms and thus enable semantic interoperability (Lutz & Klien, 2005). As the creation of rich semantic information is crucial but also difficult, tools for semi-automatic extraction of knowledge from databases and for building metadata resources are demanded to minimize the human intervention (Klien & Lutz, 2005). Furthermore, the use of multiple languages complicates matters. Depending on the data provider's nationality, metadata entries are provided in different languages. Equally, catalogue users might speak different languages. Support for information discovery in multiple languages as proposed in Nowak et al. (2005) is urgently needed in cross-nation contexts. However, concerning e.g. the area of Europe, no multilingual thesaurus that is commonly accepted for sharing a vocabulary exists (Bernard et al., 2005).

From the technological point of view, it has been widely acknowledged that research is required to address semantic interoperability issues (Egenhofer, 2002; Sheth, 1999; Sondheim et al., 1999). Developments that are under way in the Semantic Web research community are highly interesting, as they introduce new technologies to support semantic interoperability on the web. Semantic Web Services (as part of the Semantic Web research) are designed to facilitate intelligent services discovery, selection, composition and invocation in distributed environment. A combination of currently available standards-based SDIs with the developments around WSMO (web service modeling ontology), and its first reference implementation WSMX (web service modeling execution environment) could exploit the benefits of both approaches for efficient GI discovery and retrieval.

Finally, it would be worth to consider alternative developments in the area of information discovery. The approaches based on hierarchical structures like controlled vocabularies, thesauri and ontologies are justly criticized (Shirky, 2005). Instead of creating sophisticated classifications of things in the world, tagging (i.e. assigning freely chosen keywords) provides users with a collaborative way to

categorize information on the web (Golder & Huberman, 2005). This development has become known as folksonomy. Also, research in machine learning and data mining supplies some proposals to extract knowledge from semi-structured data for clustering (Fayyad et al., 1996). Clustering data from semi-structured documents has not been much explored in geodata retrieval. Clustering Geographic Markup Language (GML) documents into meaningful groups may help users to decide on the suitability of the retrieved geodata without having to bother about the huge metadata overhead.

ACKNOWLEDGEMENTS

Contributions of Prof. Werner Kuhn are gratefully acknowledged. The work presented in this paper has been supported by the ALFA-project eduGI.LA (grant **II-0426-A-FI**).

BIBLIOGRAPHY

Bernard, L., Kanellopoulos, I., Annoni, A., & Smits, P. (2005). The European geoportal—one step towards the establishment of a European Spatial Data Infrastructure. Computers, Environment and Urban Systems, 29, 15-31.

Bernstein, A., & Klein, M. (2002, June 9-12, 2002). Towards High-Precision Service Retrieval. Paper presented at the First International Semantic Web Conference (ISWC 2002), Sardinia, Italy.

Cantán, O., Nogueras-Iso, J., Torres, M. P., Zarazaga-Soria, F. J., & Lacasta, J. (2003). On the Problem of Finding the Geographic Data We Are Looking For. Paper presented at the 9th EC GI & GIS Workshop: ESDI Serving the User, A Coruña, Spain.

Craglia, M., & Evmorfopoulou, K. (2000). Comparative Evaluation of On-Line Metadata Services and User Feedback (Deliverable No. 2, INFO2000 Project: PUB1108-MADAME).

Egenhofer, M. (2002). Toward the Semantic Geospatial Web. Paper presented at the 10th ACM International Symposium on Advances in Geographic Information Systems (ACM-GIS), McLean, VA.

Fayyad, U., Piatetski-Shapiro, G., Smith, P., & Uthurusamy, R. (Eds.). (1996). Advances in Knowledge Discovery and Data Mining: MIT Press, Cambridge, MA.

Golder, S., & Huberman, B. A. (2005). The Structure of Collaborative Tagging Systems (Technical Report): Information Dynamics Lab, HP Labs.

ISO/TC-211. (2003). ISO 19115:2003. Geographic information - Metadata: International Organization for Standardization & OpenGIS Consortium.

ISO/TC-211, (2005). ISO 19119:2005. Geographic information - Services: International Organization for Standardization & OpenGIS Consortium.

Klien, E., & Lutz, M. (2005). The Role of Spatial Relations in Automating the Semantic Annotation of Geodata. Paper presented at the Conference on Spatial Information Theory (COSIT'05), Ellicottville, NY, USA.

Lutz, M., & Klien, E. (2005). Ontology-Based Retrieval of Geographic Information. International Journal of Geographical Information Science (IJGIS), forthcoming.

Manso, M. A., Nogueras-Iso, J., Bernabe, M. A., & F.J., Z.-S. (2004). Automatic Metadata Extraction from Geographic Information. Paper presented at the 7th Conference on Geographic Information Science (AGILE 2004), Heraklion, Greece.

Nielsen, J. (1999). Designing Web Usability: New Riders Publishing.

Nowak, J., Nogueras-Iso, J., & Peedell, S. (2005, 29th June - 1st July 2005). Issues of Multilinguality in Creating a European SDI - The Perspective for Spatial Data Interoperability. Paper presented at the 11th EC-GI & GIS Workshop, ESDI: Setting the Framework, Alghero, Sardinia.

OGC. (2005). OpenGIS® Catalogue Service Implementation Specification (Version 2.0.1) (OGC Implementation Specification No. 04-021r3): OpenGIS Consortium.

Richardson, R., & Smeaton, A. F. (1995). Using WordNet in a Knowledge-based Approach to Information Retrieval (Technical Report CA-0395). Dublin, Ireland: Dublin City University.

Sheth, A. P. (1999). Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics. In M. F. Goodchild, M. Egenhofer, R. Fegeas & C. A. Kottman (Eds.), Interoperating Geographic Information Systems (pp. 5-30): Kluwer.

Shirky, C. (2005). Ontology is Overrated: Categories, Links, and Tags. Retrieved 10/10, 2005, from http://www.shirky.com/writings/ontology_overrated.html

Shneiderman, B., & Plaisant, C. (2004). Designing the User Interface: Strategies for Effective Human-Computer Interaction (4 ed.): Addison Wesley.

Sondheim, M., Gardels, K., & Buehler, K. (Eds.). (1999). GIS Interoperability. New York: John Wiley & Sons.

Wachowicz, M., Vullings, W., Bulens, J., Groot, H. d., & Broek, M. v. d. (2005). Uncovering the Main Elements of Geo-Web Usability. Paper presented at the AGILE 2005 - 8th Conference on Geographic Information Science, Lisboa, Portugal.