(Draft) Proceedings of AGILE 2005: 8th Conference on Geographic Information Science. 2005, p. 359-366.

CatServer: A Server of GATOS

R. Tolosana-Calasanz, D. Portolés-Rodríguez, J. Nogueras-Iso,
P.R. Muro-Medrano, F.J. Zarazaga-Soria
Universidad de Zaragoza
Zaragoza, Spain
[rafaelt, dporto, jnog, prmuro, javy]@unizar.es

SUMMARY

This paper will present CatServer: a server of GATOS (the Spanish word GATOS means cats) a functional kernel which provides catalogue services for any XML-coded metadata. CatServer has been used for the creation of geographical dataset catalogues, catalogues of services, geocoders, Web Feature Servers and gazetteers. Nonetheless, in order to do that, we have defined GATOS as Generic dAtasets, Terms Or Services which can be described using metadata. That is to say, GATOS are the abstraction of the metadata descriptions which describe datasets, terms or services in the spatial data infrastructures context. Additionally, several techniques borrowed from the information retrieval domain have been applied in its design. These techniques provide efficiency and accuracy on query services: the Extended Boolean Model allows the system to establish a ranking among query results and the Inverted Indexes provide efficiency.

KEYWORDS: *catalogue, catalogue services, metadata, Dublin Core, information retrieval system, spatial data infrastructure, XML.*

INTRODUCTION

Most commonly defined as "structured data about data" or "data which describes attributes of a resource" or, more simply, "information about data", the concept of metadata is not new: map legends, library catalogue cards and business cards are everyday examples. Basically, metadata offer a description of the content, the quality, the condition, the authorship, and any other characteristics of the resources. It also provides for a standardised representation of information. That is, similar to a bibliographical record or a map legend, they provide a common set of terminology to define the resources or the data. Metadata constitute the mechanism to characterise data and services in order to enable other users or applications to make use of such data and services. Metadata records, each of them describing a specific resource, are grouped into catalogues thus providing the users with the possibility of finding the resources of their interest. Therefore, these catalogues are the tool to put in touch consumers with information and services providers.

Metadata cataloguing systems usually support (recognise) three forms of metadata (Nebert, 2001): the implementation form (within a database or storage system), the export or encoding format (a machine-readable form designed for the transfer of metadata between computers), and the presentation form (a human readable format). There is a general consensus about the use of XML (eXtensible Markup Language) (Bray, 2000) for the last two forms. First of all, it includes a capable markup language with structural rules enforced through a control file, in the form of a DTD (Document Type Definition) or an XML-Schema (an enhanced version of DTD defined in (Thompson, 2001)). Organisations in charge of the edition of metadata standards publish stable versions of DTDs and XML-Schemas in order to assure the conformance of metadata descriptions in XML format. And secondly, through a companion specification (XML Style Language, or XSL (W3C, 2004)), an XML document may be used along with a style sheet to produce flexible presentations or reports of content in a way that agrees with the user requirements. Additionally, the tendency of current cataloguing systems is to interchange metadata in XML according to the specific standard required by each user on demand, that is to say, they provide different views of the same

metadata. Since the base of that information management is the use of XML, this can be easily done see (Nogueras-Iso, 2004b) for details about metadata crosswalks over XML.

In this context, we have defined GATOS as Generic dAtasets, Terms Or Services which can be described using metadata. That is to say, GATOS are the abstraction of the metadata descriptions used for describing datasets, terms or services in the context of spatial data infrastructures. The base of the metadata used for GATOS is a "common core" such as the Dublin Core standard (DCMI, 2004) which can be considered as a subset of other more specific metadata standards, such as ISO 19115 see (Zarazaga-Soria, 2003). We have also developed an object model which describes different resources: geographical datasets (via ISO 19115), services, ontologies, terms in a gazetteer, or even Web pages.

This paper will present CatServer: a server of GATOS (the Spanish word GATOS means cats) a functional kernel which provides catalogue services for any XML-coded metadata. It is being used for creating geographical dataset catalogues, catalogues of services, geocoders, Web Feature Servers and gazetteers.

THE TERM GATOS

General purpose systems usually handle information from several domains which can often use a multitude of diverse metadata descriptions. An appropriate and accurate way of storing and retrieving the system's information will lead towards a more efficient, scalable, flexible and interoperable approach on managing data. These principles have guided the design and the implementation of the system described in this paper (CatServer). CatServer's design is based on the term GATOS which, according to the previous section, is the abstraction of the metadata standards used for describing datasets, terms or services; in other words, GATOS can be defined as any generic representation of any resource in a heterogeneous information system. However, in order to obtain this generic representation, a common set of descriptive elements (called "common core") has to be established. In addition, this common core should be general enough to support a broad range of purposes and business models.

Any metadata description (GATOS), managed by CatServer, might be based on any pre-existent metadata description, but they always have to include the common core. As a result, the set of CatServer's GATOS may be classified as a metadata description hierarchy where the root node is the common core (the most basic GATOS). Besides, this hierarchy can be easily extended since it provokes no system change and can be adapted to new descriptions with little effort.

Figure 1 represents the underlying metadata model used for CatServer in an actual system. In this case, Dublin Core has been chosen as the "common core" and the GATOS extremely vary in diversity, from a web page description to a well-known standard like ISO 19115 (a crosswalk between Dublin Core and ISO 19115 is defined at (Zarazaga-Soria, 2003) and (Nogueras-Iso, 2004b), they show that there is no loss of information).

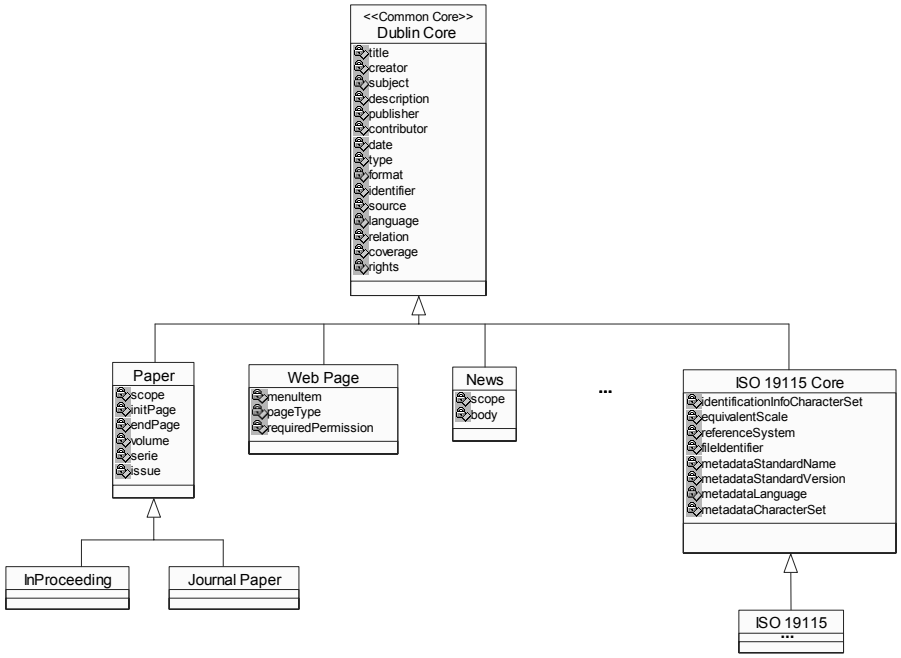

Figure 1: A GATOS model example.

CATSERVER'S FUNCTIONALITY

CatServer's design and implementation have been based on the General Catalogue Interface Model specifications by Open Geospatial Consortium (OGC) (OpenGIS Catalogue Services Specification, 2004). According to OGC, "the General Catalogue Interface Model provides a set of abstract service interfaces that support the discovery, the access, the maintenance and the organisation of catalogues of geospatial information and related resources". Consequently, our system provides management services – which allow the administrator to maintain and to organise the metadata –, discovery services – which permit querying the information – and access services – which present any results queried previously. In addition, a two-operation session component (initialise and close) is supplied for the interaction activity between the server and a client.

A brief description of the functionality might be as follows:

- Session operations. The initialise operation and the close operation ought to be the first and the last, respectively, of the operations that a client may wish to order.
- Querying the system. Query operations (discovery services) basically establish user constraints on the metadata stored in the system, in other words, a query operation tries to determine which metadata satisfy these constraints and which do not. As a result, a metadata set is obtained, but instead of sending it back as a response, the set is temporarily memorised at the server.
- Obtaining the results. Present operations (access services) have to be executed after a query operation, so that any consecutive number of the metadata which satisfy the constraints can be obtained.
- Importation and exportation of metadata. Two kinds of operations essentially exist, regarding the management functionality. In general, the import operation deals with the insertions and updates of metadata. The delete operation brings the option of removing them.

- Metadata relationship support. In some cases, certain relationships between the metadata (such as aggregations, chaining services support…) can be established in order to exploit them in the querying process. Despite its enormous interest, the techniques involved in the design of this part of the system are out of the scope of this paper, further information can be obtained at (Nogueras-Iso, 2004d).
- User control. CatServer has a user control which is inspired on the CORBA interface defined by OGC (OpenGIS Catalogue Services Specification, 2004) and which accomplishes two functional tasks. On one hand, it manages a significant security task by checking whether a client has the appropriate rights to execute the operation requested. On the other hand, every client has to create a session in order to make any request. These sessions can be exploited to store some information about the client (such as the answer obtained when carrying out a query, which is not sent back but kept at the server).
- The operation control. It was created with the aim of registering any operation that the catalogue executes. This information may be particularly useful in the near future in order to study the preferences of a user or the preferences of a group of users. As a result of this study several optimisations could be done.
- Multilingual support. A basic multilingual support is supplied since any metadata can be imported to the system in several languages and those multilingual versions are actually considered as the same metadata record. This multilingual model is explained in (Nogueras-Iso, 2004c).
- Ranking and sorting results. The results which can be obtained from a query are sorted by relevance. However, it is possible to specify other sorting criteria such as the metadata creation date.
- Disambiguation features. In the near future, a module of semantic disambiguation will be added to the system. It helps eliminate the ambiguity caused by some semantic relationships. Further information can be obtained at (Nogueras-Iso, 2004a).

Special features

Our information retrieval kernel is based on the Extended Boolean Model (Baeza-Yates, 1999) and therefore, the simplicity and the elegance of the Simple Boolean Model is combined with the slightly more sophisticated ranking of results supplied by the Extended Model. Below, the information retrieval process and some techniques applied on it will be discussed deeply. However, before that, the most important querying requirements must be enumerated.

The Discovery services permit searching metadata by means of certain constraints that have to be introduced as parameters on the operations. Additionally, the search requirements establish that those constraints will not be executed over free text but over concrete XML tags. CatServer, however, has been designed to fulfil our most important necessity: to store huge amounts of information and to be sufficiently efficient in response time. At present, it supports up to one million metadata while it is expected to be storing up to tow hundred million metadata by the near future. Besides, its response time is currently between one and five seconds approximately (depending on the amount of metadata).

Essentially, two characteristics of design have been developed in order to achieve this challenging need. Firstly, metadata are directly stored in XML at CatServer. This modus operandi is significantly different from others which convert the XML into a persistent object model. The great advantages of the adopted approach are its retrieving speed (since it *only* has to retrieve the XML) and its independence from the metadata standard; otherwise, as it happens with the persistent object model approach, new standards may mean code rewriting to be supported.

Secondly, the file structure Inverted Index (Baeza-Yates, 1999) was chosen and adapted to speed up the queries. This structure could be defined as a sequence of (key, pointer) pairs where each pointer

points to a record in a database which contains the key value in some particular field. The index is sorted on the key values to allow rapid searching for a particular key value, using e.g. binary search. The index is "inverted" in the sense that the key value is used to find the record rather than the other way round. For databases in which the records may be searched based on more than one field, multiple indexes may be created that are sorted on those keys.

Our index structure is slightly different. It consists of a pair (key, array) where the key has the same meaning but we have an array instead of a pointer to a register. The array is a metadata identifier array which represents those metadata that contain the word in a specific XML tag. The index structure has been implemented by means of a relational database table. We normally build an Inverted Index for every XML tag for which the clients need to search. In figure 2 two Inverted Indexes have been built over the Dublin Core tags title and subject.

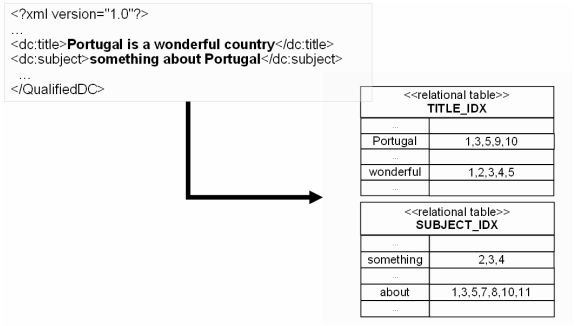

Figure 2: XML tags and Inverted Index implementation correspondence

Once the indexes are built, the system can retrieve the information: the tag name, which determines the index to examine, and the key are only needed. Consider the query represented in Figure 3 which tries to retrieve those metadata whose tag subject contains *about* or whose tag title contains *Portugal* and *wonderful* (subject LIKE about OR (title LIKE Portugal AND title LIKE wonderful)). Thus, CatServer would obtain three arrays of metadata identifiers: one for *Portugal*, another one for *wonderful* and another for *about*. The next step in the process is to combine these arrays as sets of metadata. The AND implies an intersection operation between the *Portugal* array and the *wonderful* array. The OR implies a union operation between the *about* array and the subset obtained in the previous step.

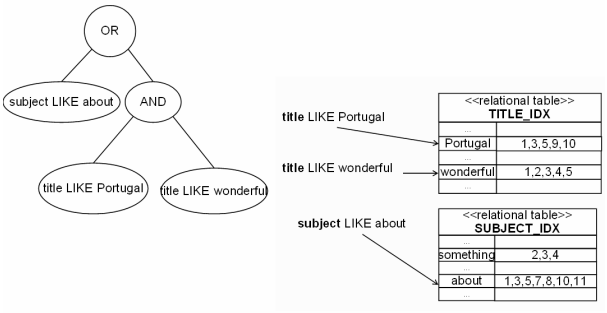

Figure 3: Example on the querying process

Evidently, not all the results are equally important. As mentioned above, at the beginning of this section, the ranking process is based on the Extended Boolean Model; therefore, the subset of metadata is in fact a list of metadata ordered by relevance. Following with the example, it seems that those metadata which satisfy both OR conditions are actually more important than those which only satisfy one of them. For that reason, they appear before on the list.

The initial value of the metadata is also a way for establishing differences on the metadata depending on which organisation created them. Nevertheless, it was considered that this initial weight should depend on the XML tag as well, since some tags could actually be more relevant than others.

Additionally to what has been explained, not only may the answer be ranked or sorted by relevance, but also by a certain criterion such as the "data creation date". Because of this requirement and because of our efficiency reasons, CatServer precalculates and stores the results of those criteria which need sorting. This metadata list contains all the metadata stored in the system sorted by the respective criterion. Hence, the last step on a querying process in which it is specified a sorting criterion does not rank the answer by relevance. Once the query response is obtained, the system retrieves the ordered list associated to this criterion and sorts the query response according to that list.

PUTTING CATSERVER TO WORK

CatServer has been designed to be a metadata retrieval kernel which is likely to be integrated in the components of a SDI. Since the retrieval and query requirements are different among SDI components, building a system of CatServer's features is not easy. Nonetheless, our GATOS hierarchy approach also simplifies this retrieval kernel development. As long as a metadata standard can be inserted in our hierarchy, and consequently it can be managed by CatServer, the metadata heterogeneity is reduced. This fact simplifies the design, the implementation and the maintenance.

Up to date CatServer has been used in several Metadata Catalogues, Services Catalogue, Gazetteers, Geocoders and portals, as the following enumeration shows:

- According to OGC a catalogue is a component that supports the ability to publish and to search collections of metadata for data, services and related information objects. If this definition is restricted to geographical metadata, we have a geographical metadata catalogue which stores descriptions of the geographical information in the SDI. Two metadata catalogues which have been built by means of CatServer's technology are the geographical catalogue of the Spanish SDI accessible at http://www.idee.es, which stores about thirty thousand metadata, and the geographical catalogue of the Zaragoza City Council accessible at http://idezar.unizar.es with about two hundred metadata.
- A services catalogue stores descriptions of services supplied somewhere. The Spanish SDI services catalogue (http://www.idee.es) uses CatServer as well.
- A Gazetteer service is a network-accessible service that retrieves the known geometries, for one or more features, given their associated well-known feature identifiers (text strings). The Spanish SDI gazetteer (accessible at http://www.idee.es) is built upon CatServer's technology and manages about one million metadata.
- A Geocoder Service is a network-accessible service that transforms a description of a feature location, such as a place name, street address or postal code into a normalised description of the location, which includes a coordinate geometry. An example of a Geocoder which currently has CatServer's is the Zaragoza City Council Geocoder (http://idezar.unizar.es) which stores about six thousand metadata.
- A Portal is a web site on the Internet which people use to search. Sometimes portals access to database repositories to obtain the information. That is the case of the Spanish SDI portal which (http://www.idee.es) provides information about news, web pages and organisations by means of CatServer.

Additionally, CatServer will be used in the development of new components such as a Web Ontology Server (currently still in prototype), a catalogue of images and other components in which retrieval services are the base of the functionality.

CONCLUSIONS AND FUTURE WORK

This paper has presented CatServer (a server of GATOS), a functional kernel which provides catalogue services for any XML-coded metadata. It is a metadata server specialised in the Spatial Data Infrastructure context. CatServer's technology is currently in execution at the most important Spanish SDI nodes with notable success, since it supports up to one million metadata and performs accurate queries in reasonable response times.

This success is achieved because of two main design features: the GATOS hierarchy approach and several techniques borrowed from the information retrieval field. By means of the GATOS approach, CatServer has been able to satisfy important information system characteristics such as efficiency, scalability, flexibility and interoperability. Thus, the metadata description model is simple and elegant and can be easily extended. Besides, the information retrieval techniques applied provide efficiency and accuracy on query services: the Inverted Indexes provide efficient metadata retrieval and the Extended Boolean Model allows the system to establish a ranking among the query results.

As a consequence, it has also been shown how CatServer has been used as a metadata retrieval kernel in several SDI components: Gazetteers, Geocoders, Metadata, Services catalogues and Portals have its technology integrated with high success. In addition, it is expected to be used in the development of other components in which retrieval services are the base of the functionality.

Future research work may be the improvement of the data structures in order to increase the metadata number that the system can handle. We are currently considering an efficient retrieval information system design able to manage enormous amounts of metadata (up to two hundred million metadata) which can also be interrelated. Those relationships between metadata may be of any semantic type.

ACKNOWLEDGEMENTS

The basic technology of this work has been partially supported by the Spanish Ministry of Science and Technology through the project TIC2003-09365-C02-01 from the National Plan for Scientific Research, Development and Technology Innovation.

BIBLIOGRAPHY

Baeza-Yates R., Ribeiro-Neto B., 1999: Modern Information Retrieval. Addison Wesley. ISBN 0-201-39829-X

Bray T., Paoli J., Sperberg-McQueen C. M., Maler E., 2000: Extensible Markup Language (XML) 1.0 (Second Edition). W3C. W3C Recommendation. http://www.w3.org/TR/2000/REC-xml-20001006

Core Metadata Inititiative (DCMI), 2004: Homepage of the Dublin Core Metadata Initiative. http://www.dublincore.org

Nebert D, 2001.: Developing Spatial Data Infrastructures: The SDI Cookbook v.1.1. Global Spatial Data Infrastructure. http://www.gsdi.org

Nogueras-Iso J., Lacasta J., Bañares J.A., Muro-Medrano P.R., P., Zarazaga-Soria F.J., 2004a: Exploiting disambiguated thesauri for information retrieval in metadata catalogs. Lecture Notes in Artificial Intelligence Volume 3040, 322-333.

Nogueras-Iso J., Zarazaga-Soria F.J., Lacasta J., Béjar R., Muro-Medrano P. R. 2004b: Metadata Standard Interoperability: Application in the Geographic Information Domain. Computers, Environment and Urban Systems, 28, 611-634.

Nogueras-Iso J., Zarazaga-Soria F.J., Lacasta J., Tolosana R., Muro-Medrano P.R., 2004c: Improving multilingual catalog search services by means of multilingual thesaurus disambiguation. Proc. of the 10th EC-GI & GIS Workshop ESDI: The State of the Art.

Nogueras-Iso J., Zarazaga-Soria F.J., Muro-Medrano P.R., 2004d: Management of nested collections of resources in Spatial Data Infrastructures. Proc. of the 1st International Workshop on Geographic Information Management (GIM'04) 878-882.

OpenGIS Catalogue Services Specification: OGC 04-02lr2, Version 2.0. 11-05-2004. Open Geospatial Consortium (OGC). Homepage of the OGC http://www.opengeospatial.org

Thompson H. S., Beech D., Maloney M., Mendelsohn N.: XML Schema Part 1: Structures. W3C, 2001. W3C Recommendation. 2 May 2001. http://www.w3.org/TR/2001/REC-xmlschema-1-20010502/

W3C: The Extensible Stylesheet Language Family (XSL), 2004. http://www.w3.org/Style/XSL/

Zarazaga-Soria F.J., Nogueras-Iso J., Ford M.: Mapping between Dublin Core and ISO 19115, "Geographic Information – Metadata". CEN/ISSS Workshop - Metadata for Multimedia Information - Dublin Core. September, 2003. Number 14857.