9th EC GI & GIS Workshop, ESDI Serving the User, A Coruña, Spain, 25-27 June 2003

Integrating geographic and non-geographic data search services using metadata crosswalks

F.J. Zarazaga-Soria, M.P. Torres, J. Nogueras-Iso, J. Lacasta, O. Cantán
Computer Science and Systems Engineering Department
University of Zaragoza
Maria De Luna 1, 50018 Zaragoza (Spain)
{javy, mapitb, jnog, jlacasta, ocantan}@unizar.es

Abstract. There are many scenarios that require the combination of different types of metadata. For instance, let us imagine the case of three different metadata-databases describing the elements from a library (books, reports and other kinds of documents), events (movies, theatres, recitals, etc) and geographic data (maps, satellite images, etc) respectively. These databases could be integrated into systems that provide specialized high-level services such as tourist information (events and publications can be linked with data for travelling to a tourist destination) or cultural-information (publications can be linked to an event, and it could be useful to provide maps for accessing the places where the event occurs). However, the metadata stored in the different databases are not compatible each other because they must conform to a specific domain standard (e.g. USMARC for library, Dublin Core for events and ISO 19115 for geographic data). One flexible and efficient way for developing these high-level services could be the use of metadata-crosswalks to unify the metadata-access (search and retrieval) methods. The implementation of these crosswalks should be done using emergent technologies like XML and XSL. In the aforementioned scenario, the use of these technologies would provide the mechanism for employing the same methods to query the three databases. This paper will present a metadata crosswalk-based approach for integrating services to search geographic and non-geographic data in a homogeneous way.

1. INTRODUCTION

Nowadays, information is very heterogeneous and it is widely distributed. Thanks to Internet, information and resources are accessible to anybody from anywhere. But such amount of data and services involve some problems too. Among these problems, interoperability is perhaps one of the most important. By interoperability, it is meant the ability to develop conventions so that data exchange and integration become possible. As a special kind of interoperability, semantic interoperability is the agreement about content description standards. Nowadays, there are an increasing number of initiatives that aim at solving the problem of semantic interoperability on the Web.

For this reason, good descriptions of metadata and resources are necessary. Therefore, the interest in resource descriptions that are easy to create and that almost anyone can understand has been growing steadily. Metadata plays an important role in this field. 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 and everyday examples can be found at map legends, library catalogue cards or business cards. Basically, metadata offers description of the content, quality, condition, authorship, and any other characteristics of a resource. It also provides standardized representation of information. That is to say, similar to a bibliographical record or map legend, it provides a common set of terminology to define the resource or data.

Metadata records, each one describing a specific resource, are grouped into catalogues that provide users with the possibility of identifying the resources of their interest. In order to minimize the cost of time for the creation and maintenance of metadata and to maximize its usefulness to the wider audience of users, it should be desirable to use a unique metadata standard in storage labours and provide automated views of metadata in other related standards. According to this philosophy, the tendency of the current cataloguing systems is to interchange metadata in XML that conforms to a specific standard on user demand, that is to say, providing different views of the same metadata.

Let us imagine a scenario where three different metadata-databases store meta-information that describes the elements from a library (books, reports and other kinds of documents), events (movies, theatres, recitals, etc) and geographic data (maps, satellite images, etc) respectively. These databases can be used for providing specialized high-level services such as tourism-information (events and publications can be linked with data for travelling to a tourist destination) or cultural-information (publications can be linked to an event, and it could be useful to provide maps for accessing to the places where the event occurs). The problem is that standards used in each metadata-database should conform to a specific domain and it will be necessary to unify the metadata-access (search and retrieval) methods. Figure 1 displays the scenario described above and the different databases that must be integrated.

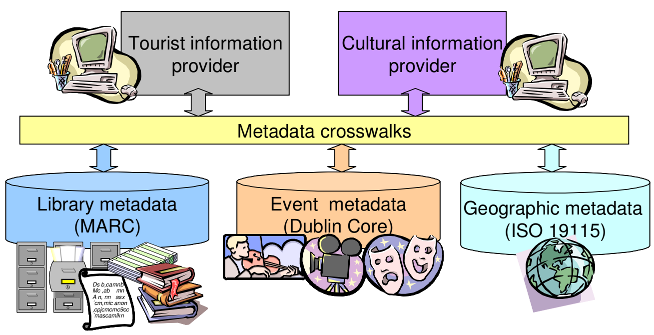

Figure 1: Use case of metadata interoperability

Library metadata database in Figure 1 specifies the use of MARC standards. MARC is the acronym for MAchine-Readable Cataloguing and is one of the most widely used standards in the library application domain. It defines a data format which emerged from a Library of Congress led initiative begun thirty years ago. MARC became USMARC in the 1980s and MARC 21 in the late 1990s. It provides the mechanism by which computers exchange, use and interpret bibliographic information and its data elements make up the foundation of most library catalogs used today [1].

As far as geographic information domain is concerned, there are two main metadata standards: CSDGM [2, 3] and ISO 19115 [4]. CSDGM stands for "Content Standard for Digital Geospatial Metadata". It was carried out in 1994 by the Federal Geographic Data Committee (FGDC) of the United States to give support for the construction of the National Spatial Data Infrastructure. And although it is a national standard, it is the oldest one and has been incorporated into many GIS tools and networks (e.g. the Clearinghouse project), thus becoming the most widely used in GIS world (e.g. adopted in countries like

South Africa or Canada). The other big standard for geographic information metadata is ISO 19115. The organization responsible for this standard is the International Standard Organization (ISO) who created in 1992 the committee 211 (ISO/TC 211) with responsibilities in "geomatics". It has the rank of international standard and has been recently released (May 2003). Therefore, the current trend in Geographic Information world is the use of ISO 19115, instead of CSDGM. The system shown in Figure 1 displays ISO 19115 as the standard for geographic metadata.

Finally, event metadata database specifies the use of Dublin Core standard [5]. Dublin Core is a metadata standard of general outreach which has been also recognized as ISO standard, the ISO 15836 [6]. The scope of the Dublin Core was specifically designed to provide a metadata vocabulary of "core" properties able to provide basic descriptive information about any kind of resource, regardless of the media format, area of specialization or cultural origin. In the scenario depicted in Figure 1 Dublin Core is used to describe events but it is also widely used for e-government or even to describe minimally a geographic resource.

In order to make these metadata databases interoperable, it will be necessary to provide the description of data and resources in different standards (and if possible, at low cost). There are two main approaches to handle the semantic interoperability problem between metadata standards: solutions that are based on the use of ontologies (i.e. establishing or inferring relationships between the metadata vocabularies employed by the different metadata standards); and the creation of specific crosswalks for one-to-one mapping.

Ontology is defined as an explanation of some shared vocabulary or conceptualization of a specific subject matter, and it seems to be an adequate methodology that helps to define a common ground between different information communities. Furthermore, approaches that aim at solving the problem of semantic interoperability on the Web by means of ontologies are closely related to a new conception of the Web: the Semantic Web. According to [7][8], "the Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation".

These approaches offer very flexible solutions for interoperability. However, this ambitious aim of flexibility may also imply a lack of accuracy in the mappings performed. The ontology based solutions do not consider the local structural constrains imposed by the different specific domains, e.g. parent/child relationships; cardinality/occurrence constraints; datatyping, enumeration and formatting constraints on the element values. As stated in [9]: "the wider the targeted scope of interoperability, the more difficult it is to achieve accurate, precise mappings". For a small set of metadata standards, whose syntax and semantics are relatively fixed and constrained, hardwired crosswalks establishing the mapping between metadata terms (from specific standards) may result more adequate than ontology-based solutions. That is precisely the case of the scenario presented as example.

The rest of the article is structured as follows. Next section deals with the construction of metadata crosswalks. First of all, it presents the related work done in this field, and secondly, it summarizes the process and the technology used. Section three comes back to the example presented above and explains the means for providing interoperability in such scenarios. Finally, this work ends with a section of conclusions.

2. CONSTRUCTING METADATA CROSSWALKS

2.1 Related Work

There is a big experience in developing mappings among several standards and different domains. Interesting collections of links to metadata-crosswalk initiatives can be found at http://www.sinica.edu.tw/~metadata/tool/mapping-foreign.html

and http://www.ukoln.ac.uk/metadata/interoperability/. There, it is possible to find mappings from MARC 21 to Dublin Core, Dublin Core to USMARC, Dublin Core to EAD/GILS/USMARC, Dublin Core to FINMARC/GILS, Dublin Core to IAFA/ROADS templates, Dublin Core to UNIMARC, FDGC to GCMD DIF, FGDC to USMARC, and others.

The Canadian Heritage Information Network (CHIN, http://www.chin.gc.ca/) [10] offers some links to crosswalks that may be of use to museums. Some examples of the links offered could be the "Crosswalk of Metadata Element Sets for Art, Architecture, and Cultural Heritage Information and Online Resources" (developed by the Getty Research Institute, its mapped standards include Categories for the Description of Works of Art, VRA Core Categories, Dublin Core, Object ID, the CIMI Access Points, the Guide to the Description of Architectural Drawings, as well as library and archival standards), or the Mapping from CHIN Natural Sciences Data Dictionary to Darwin Core (CHIN has completed a mapping between the Darwin Core and the CHIN Natural Sciences Data Dictionary so museums following the CHIN Natural Sciences Data Dictionary could use the same or similar mapping to Darwin Core).

As long as the geographic information metadata is concerned, the MADAME project (http://www.shef.ac.uk/~scgisa/MADAMENew/faq.html) developed a correspondence between Dublin Core and ISO19115.3. This correspondence, which can also be found at the ETeMII project (http://wwwlmu.jrc.it/etemii/index.html) document [11], offers a table with the correspondence between the Dublin Core sections and the ISO 19115.3 sections, but it does not offer any automatic or semi-automatic tool for transforming from one to the other. It also provides a correspondence between prENV 12657 and ISO TC 211 /CD 19115.3 (http://www.shef.ac.uk/~scgisa/MADAMENew/cen2iso.pdf) with similar limitations.

The Canadian Geospatial Data Infrastructure has also developed a crosswalk between ISO19115 and FGDC standard (see [12], [13]). The discovery portal of this infrastructure (GeoConnections Discovery Portal at http://geoconnections.ca) offers data products catalogued in accordance with the FGDC CSDGM standard but plans to support the new ISO19115 in future versions.

Additionally, the DGIWG (Digital Geographic Information Working Group) Metadata Work Program, supported by NIMA (National Imagery and Mapping Agency of United States), offers a crosswalk between ISO19115 and FGDC standard (http://metadata.dgiwg.org/ISO19115/related.htm) too. This program is taking a leading role in developing an implementation model and XML schema of the ISO 19115 metadata standard (officially known as ISO 19139) and provides a Metadata Development Efforts Website (http://metadata.dgiwg.org/) to coordinate the metadata standardization efforts of several organizations.

On the other hand, the own FGDC organization provides a mapping between FGDC standard and Dublin Core. It is available at http://geology.usgs.gov/tools/metadata/tools/doc/dublin.html. And there, it is explained how the *mp* tool (parser of formal metadata provided by FGDC) generates an HTML output, where FGDC elements are mapped to Dublin Core elements in the META tags of the HEAD section. The intended use of META tags is to divulgate the content of a Web page, thus making this meta-information visible to search engines.

Inside the project "Cooperative Online Research Catalog (CORC)" a converter among FGDC, Dublin Core and MARC21 standards was developed as one of its goals [14][15]. One of the motivations of this work was the unsuccessful results (on average) obtained from queries directed at nodes of the FGDC Clearinghouse [16]. Therefore, it was proposed to convert FGDC metadata into more widely used metadata standards for inclusion in systems other than the FGDC Clearinghouse.

Most specifically within the context of environmental geographic information, a mapping between ISO and GELOS has been built inside the project "ETC/CDS (EIONET): European Topic Centre on Catalogue of Data Sources" (http://eionet.eu.int/). Another work is the mapping between UDK-metadata standard and ISO. This mapping has been developed inside the project "UDK (Umwelt Data Katalog)" (http://www.umweltdatenkatalog.de/), German Environmental data catalog.

Most of these works do not include any other result apart from the table that maps the relationships and equivalencies among the standards. In some cases, any kinds of tools for automatic or semiautomatic translation are included. And almost no-one offers details about the process followed. In this sense, there are two interesting works that manage this problem. In [17] some of the common misalignments in creating crosswalks are presented. The other interesting work is [18]. It provides many of the key issues involved in crosswalk development and identifies those areas in which harmonization can contribute. As the paper explains, its main contribution is the delineation of the general issues involved in the harmonization of metadata standards and in the development of crosswalks between related metadata standards. Many concepts and ideas presented in it have been used as a base for the development of the work presented in this paper.

2.2 Mapping Creation Process

In order to maintain this interoperability across related metadata standards, it is necessary the creation of software systems able "to speak several metadata dialects", that is to say, systems that provide *crosswalks* between metadata standards. According to the Dublin Core Metadata Glossary [19]: "A crosswalk is a table that maps the relationships and equivalencies between two or more metadata formats. Crosswalks or metadata mapping support the ability of search engines to search effectively across heterogeneous databases, i.e. crosswalks help promote interoperability".

The construction of crosswalks between standards is much more than the use of a series of programming technologies. A crosswalk specifies the mapping between two related standards, thus enabling communities that use one standard to access the content of elements defined in another one. Unfortunately, the construction of crosswalks constitutes a difficult and error-prone task that requires deep knowledge and vast experience with the standards. The obtainment of the knowledge required to construct a crosswalk is particularly problematic since each metadata standard has been developed frequently in an independent form; and therefore, different terminology, specialized methods and processes are used. Moreover, the maintenance of crosswalks between metadata standards which are not stable and subject to changes is even more problematic due to the additional requirement of adjusting crosswalks to historical versions. The process followed for creating the mapping proposed in this work is structured in for main steps:

- Harmonization: This phase aims at obtaining a formal and homogeneous specification of both standards.
- Semantic mapping: In order to determine the semantic correspondence of elements between the standards of metadata a deep knowledge of the origin and destiny metadata standards is required. As result of this phase, a mapping table is created.
- Additional rules for metadata conversion. Apart from the mapping table, it should be necessary to provide additional metadata conversion rules in order to solve problems such as different level of hierarchy, data type conversions, etc. .
- Mapping implementation: The last objective of the process is to obtain a completely automated crosswalk by means of the application of some type of tool. In this way, maintaining only one set of metadata, searches and views can be provided according to the different families from metadata.

Details about these steps can be found in [20]. Taking into account that the metadata standards presented in the introduction section use XML as exchange and presentation format, it has been considered that the most suitable technology to carry out the implementation of crosswalks is by means of XSL (eXtensible Stylesheet Language [21]), whose purpose is precisely the manipulation and transformation of XML. XSL is a language for expressing style sheets that integrates two related languages: a transformation language (XSL Transformations or XSLT); and a formatting language (XSL Formatting Objects) of XML documents, which is comparable to the language CSS (Cascading Style Sheets) for HTML pages. The transformation language (XSLT) provides elements that define rules to transform an XML-document into another XML-document. This second document can use the same set of elements that the original document (it is associated to the same DTD or XML-Schema) or can use a completely different set of elements.

3. AN APPROACH FOR THE INTEROPERABILITY

Let us come back to the example presented in the introduction. It is clear now that we need crosswalks (created according to the process sketched in previous section) for translating from one standard to another, in both senses. But having a set of isolated crosswalks is not enough. In order to provide interoperability in this scenario, it is necessary to have a system that facilitates the independence among the different actors involved. That is to say, such a system should hide metadata crosswalk details from information providers (cultural or tourist information providers) and databases. For instance, as times goes by, the standards will probably add some modifications. Therefore, in order to maintain the coherence and correctness between the modified metadata standard and the rest of the standards, it will be necessary to revise and maintain our crosswalks up-to-date. However, such a change only minds the people in charge of the crosswalks where this modified standard takes part. People who are browsing tourist or cultural information do not care whether something has happened with standards. Crosswalks managers are the unique responsible for updating them whenever it is needed.

This system, which facilitates the integration and coordination of metadata crosswalks, has been titled “metadata crosswalk broker”. It consists of two main components: a repository where the different crosswalks are collected; and, a mediator component that makes the translation from one standard to another using a concrete crosswalk. In general, this last component acts as an intermediary. It receives the query from the client (using restrictions over the metadata elements of a particular standard); it selects the appropriate crosswalk of the repository; it applies this crosswalk to transform the query to the equivalent one in another metadata standard; and finally, it performs the query against the intended database. Then, the database will return the results to the metadata crosswalk broker; and this mediator component will be in charge of translating them again to the standard required by the client.

In the referred example, a tourist information provider, which manages metadata in Dublin Core, could perform queries to the system with restrictions over Dublin Core metadata elements, whereas the cultural-information provider could query the system using MARC. The metadata-crosswalk broker will provide the necessary crosswalks for the transformation among Dublin Core, MARC and ISO 19115. Figure 2 displays the interactions between the different components when a user makes a question.

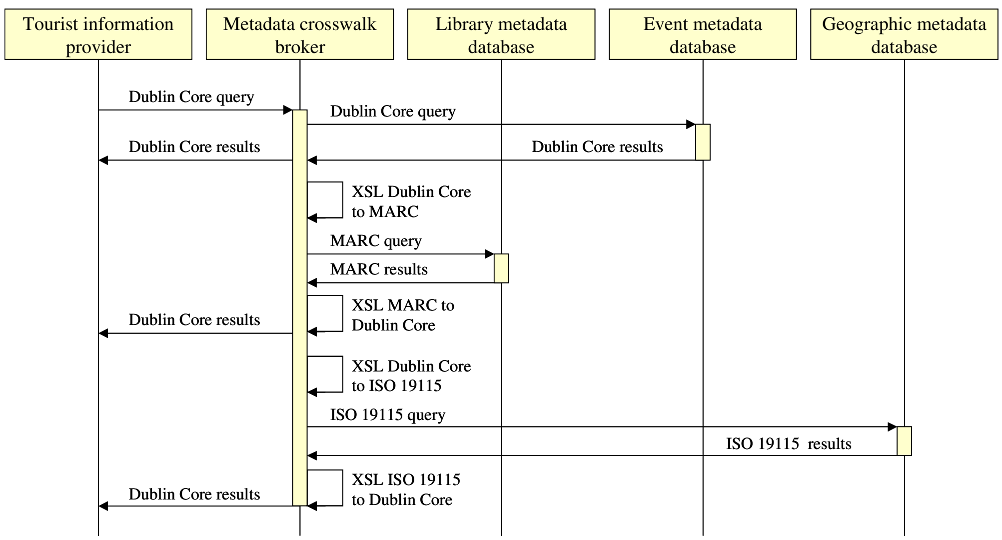

Figure 2: Crosswalks applied for the tourist information provider use case

The tourist information provider makes a query to the system in Dublin Core. Then, the query is translated to the standard used by the database in case the standards are different. The first database asked is the Event metadata database, which has all metadata in Dublin Core. Therefore, in this first search no translation must be made, and when the results are found, they are sent to the user by the metadata crosswalk level. The second database asked is the Library metadata database, which has the metadata in MARC. Hence, the metadata crosswalk broker must translate the query from Dublin Core to MARC, and then, it is sent to the Library database. When the metadata crosswalk broker receives the results, they are translated from MARC to Dublin Core and sent to the user. And finally, the third database asked is the Geographic metadata database, which has the metadata in accordance with ISO 19115. Again, at the metadata crosswalk broker the query must be translated from Dublin Core to ISO 19115 and sent to the database. When the results are ready, they must be translated from ISO 19115 to Dublin Core and they are sent to the user. If the cultural-information provider needed to query the system, a similar process would be launched to perform the necessary translations between MARC and the rest of standards.

4. CONCLUSIONS

This work has proposed the construction of a series of utilities that facilitate access to systems that use and manage metadata according to different standards. These utilities are based, mainly, on the use of crosswalks. On one hand, thanks to them, a user will be able to access to multiple metadata databases. And on the other hand, a company that maintains metadata in accordance with in a concrete standard (maybe an own standard) will be able to provide other views of these metadata without rewriting all the information, just implementing some crosswalks and applying them automatically.

The tourist information scenario has proved itself to be a valid example that could be extended to other domains. Internet surfers who plan to travel abroad can use Dublin Core in order to obtain basic guidance about the places they are looking for. Full accessibility to the culture and its services still requires deep knowledge of the local vocabulary and environment. But a set of simple facts found in Dublin Core metadata can bring the tourist attention to a foreign information portal, which does not necessarily manage Dublin Core metadata and which otherwise have gone unnoticed.

Due to the heterogeneity of metadata standards and the necessity for interoperability, many public and private entities are working actively in the construction of crosswalks. At the university of Zaragoza, the process for building crosswalks has been applied to develop the crosswalks between FGDC and ISO; FGDC and Dublin Core; and ISO and Dublin Core.

ACKNOWLEDGEMENTS

The basic technology of this work has been partially supported by the Spanish Ministry of Science and Technology through the projects TIC2000-1568-C03-01 from the National Plan for Scientific Research, Development and Technology Innovation and FIT-150500-2003-519 from the National Plan for Information Society; and by the Aragón Government through the project P089/2001. The work of J. Lacasta (ref. B139/2003) and O. Cantán (ref. B119/2001) has been partially supported by a grant from the Aragón Government and the European Social Fund.

REFERENCES

[1] Library of congress Network Development and MARC Standards office. http://www.loc.gov/marc/marcspa.html.
[2] "Content Standard for Digital Geospatial Metadata Workbook", version 2.0. Federal Geo-graphic Data Committee (USA), 2000.
[3] Document FGDC-STD-001-1998 Content Standard for Digital Geospatial Metadata. Meta-data Ad Hoc Working Group, Federal Geographic Data Committee (USA), 1998
[4] "Draft International Standard ISO/DIS 19115, Geographic information — Metadata". ISO/TC 211 (http://www.isotc211.org), dated September 2001.
[5] Homepage of the Dublin Core Metadata Initiative: http://www.dublincore.org ", dated May 2002.

[6] ISO 15836:2003(E). "Information and documentation – The Dublin Core metadata element set"

[7] Berners-Lee, T., Hendler, J., Lassila, O., "The Semantic Web", Scientific American, May 2001.

[8] "The semantic web", W3C Technology and Society Domain, Semantic Web Activity, http://www.w3.org/2001/sw/

[9] Hunter, J., "MetaNet - A Metadata Term Thesaurus to Enable Semantic Interoperability Between Metadata Domains". Journal of Digital information, volume 1 issue 8, Aug. 2001.

[10] Sherwood, L.E., "Standards for access to museum content: practical solutions or technical chimeras?", Computer Standards & Interfaces, vol. 20 (1998), pp 111 – 115

[11] Craglia, M., "Towards a European Approach to Metadata for Geographic Information", ETeMII project document (http://www.ec-gis.org:8080/wecgis/docs/F25765/D421-METADATA.PDF).

[12] "The Access Technical Services Manual Version 1.1" at http://www.geoconnections.ca/

[13] Teng, Y., "Use of XML for Web-Based Query Processing of Geospatial Data". Master Thesis available at http://www.cs.unb.ca/tech-reports/files/TR00_135.pdf.

[14] Chandler, A., Foley, D., Hafez, A.M., "Mapping and Converting Essential Federal Geo-graphic Data Committee (FGDC) Metadata into MARC21 and Dublin Core. Towards an Al-ternative to the FGDC Clearinghouse" D-Lib Magazine, January 2000, Volume 6 Number 1 (also at http://webdoc.sub.gwdg.de/edoc/aw/d-lib/dlib/january00/chandler/ 01chandler.html).

[15] Chandler, A., Foley, D., Hafez, A.M., "DC-7 Case Study: Mapping Essential Federal Geographic Data Committee (FGDC) Metadata Into Dublin Core", available at http://eeirc.nwrc.gov/pubs/crosswalk/dc7casestudy.htm, 4 October 1999.

[16] "Homepage of the National Geospatial Data Clearinghose: http://www.fgdc.gov/clearinghouse/clearinghouse.html ", National Spatial Data Infrastruc-ture, dated May 2002.

[17] Woodley, M., "Crosswalks: the Path to Universal Access?" Getty Research Institute, Getty Standards and Digital Resource Management. Dated May 7, 2000 http://www.getty.edu/research/institute/standards/intrometadata/2_articles/woodley/index.html

[18] St. Pierre M., LaPlant W.P., 1998. "Issues in crosswalking Content Metadata Standards". National Information Information Standards Organisation, http://www.niso.org/press/whitepapers/crsswalk.html, dated 1998.

[19] Dublin Core Metadata Glossary, Final Draft, Feb. 24, 2001. Available online at http://library.csun.edu/mwoodley/dublincoreglossary.html

[20] Lacasta, J., Nogueras-Iso, J., Torres, M.P., Zarazaga-Soria, F.J.: "Towards the Geographic Metadata Standard Interoperability". Proc. of the 6th AGILE Conference, Lyon (France), 23 – 26, April 2003

[21] The Extensible Stylesheet Language (XSL). http://www.w3.org/Style/XSL/