Community Metadata ISO 19115 Adaptor

Qiping Yan, Michael J. McMahon Jr., Sergiu Dascalu, Frederick C. Harris, Jr., Likhitha Ravi
Computer Science and Engineering Dept.
University of Nevada-Reno, Reno, Nevada 89557
qiping@aol.com, mcmahon@cse.unr.edu, dascalus@cse.unr.edu, Fred.Harris@cse.unr.edu, likhitha.ravi@gmail.com

Abstract

In today's world, geographic data plays an increasing role in many areas, including academic research, government decision making, and in people's everyday lives. As the quantity of geographic data gets larger, making full use of the data in a distributed, heterogeneous network environment, like the Internet, becomes a major issue. To better utilize and share those valuable resources, metadata standards have been developed. Metadata makes it easier to discover, explore, and share geographic data, particularly for cataloguing geographic data in clearinghouses. But one of the new problems that came about with metadata is interoperability since multiple metadata standards exist. Important geographic metadata standards include: ISO 19115, Dublin Core, CSDGM (US) and prENV 12657 (Europe). Semantically, metadata standards are distinct but, rather, they overlap and relate to each other in diverse and complex ways. In this research, we propose a mechanism to transform different geographic metadata standards to solve the interoperability issue using the new popular web services and XML/XSLT technologies.

Keywords: Geographic metadata, metadata interoperability, metadata standards, adaptors, ISO 19115.

1 Introduction

1.1 Metadata

Metadata is commonly defined as "data about data", "structured data about data", or "data which describes attributes of a resource", or even "information about data". Metadata provides information about one or more aspects of the data, such as purpose of the data, means of creation of the data, owner or author of the data, point of contact of the data, etc.

Metadata is not a new concept. Business cards and library cards are examples of metadata in our everyday lives. Metadata is often used for photographs (IPTC Schema, XMP, Exif, PLUS, etc.), videos (transcripts, text description of the scenes), and webpages (keywords, description, software used to create the page, author of the page).

Furthermore, metadata is commonly used for computer or software, rather than humans. Metadata can be stored internally, in the same file as the data, or externally, in a separate file.

1.2 Metadata Standards

As described in [4], "metadata standards are requirements which are intended to establish a common understanding of the meaning or semantics of the data, to ensure correct and proper use and interpretation of the data by its owners and users." A metadata standard is usually established by national and international standard communities like ANSI (American National Standards Institute) and ISO (International Organization for Standard).

1.3 Geographical Metadata Standards

In the geographic domain, the most common metadata standards are FGDCs Content Standard for Digital Geospatial Metadata (CSDGM) and the recently ratified ISO 19115 [12]. The CSDGM standard contains over 300 data and compound elements while the ISO 19115 has over 400 elements (divided into 14 metadata packages) in 86 classes that have 282 attributes and 56 relations. The ISO 19115 was developed by the geospatial community to address specific issues relating to both the description and the curation of spatial data [5]. The ISO 19115s abstract models are written using the UML (Unified Modeling Language). The accompanying XML schema, ISO/CD TS 19139, enables interoperable XML expression of ISO 19115 compliant metadata.

1.4 XML/XSLT

The XML (eXtensible Markup Language) is a markup language created to structure, store, and transport data. It defines a set of rules for encoding documents in a format that is readable by both machines and human readers. Its main purpose is to separate data from its presentation. The design goals of XML emphasize simplicity, generalization and us- ability over the Internet. It is a textual data format with strong unicode support for the languages of the world. XML is a very flexible markup language because users can create their own tags to structure and store their documents.

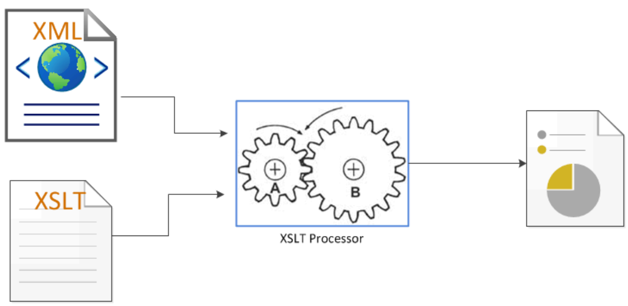

Figure 1: An XSLT processor takes one or more XML documents and one or more XSLT files as input and produces a new XML document as output

On the other hand, XSLT (eXtensible Stylesheet Language Transformation) is a language for transforming an XML document into other XML documents, or other objects. As shown in Figure 1, the XSLT template processor takes one or more XML source files and one or more XSLT style-sheet modules, and processes them with the XSLT template processing engine to produce an output document.

2 Background

In the information age, data, especially geographic data, is exploding. Many companies produce and collect geo-data in various formats from different sources, and with different types of equipment. The result of this heterogeneous geo-data creates a major problem for data sharing. Geographic metadata describes the existing geo-data. By reading the geo-metadata, users can get more information about the original dataset, like name, quality, ratio, data structure, etc. [11].

Because most geo-data is large in size, written in various (standard) formats and stored in different file formats, it is hard to directly access the original dataset. Thus, it is necessary to utilize the geo-metadata to not only describe and catalogue data, but also to discover, convert, manage, and use data in a network [9].

The more widely used geographic metadata stan- dards include CSDGM by FGDC and ISO 19115 by ISO/TC211 [4]. There are many differences between metadata standards. For example, the original version of DC (Dublin Core) defined only 15 elements known as the original set of the 15 classic metadata elements, while the ISO 19115 has more than 300 elements, organized in 86 classes with 282 attributes and 56 relations.

Metadata standards are usually presented using a structural file. This concept is useful for standards with many elements defined, but makes it hard to analyze and process the metadata. Moreover, there is not a single metadata definition language. Therefore, different standards are presented using different notations. For example, ISO 19115 uses the UML while CSDGM uses a formal file notation [13]. Finding a common way to present different metadata standards is a must in order to make it possible for computers to automatically recognize, analyze metadata, and share geo-data in different metadata standards. To this end, XML is a popular markup language that can define, present, verify, and index metadata [10].

3 Proposed Approach

3.1 Overall Solution

This effort is part of the metadata exchange project, which is, itself, part of Track II of the NSF EPSCoR funded project “Collaborative Research: Cyber-infrastructure Developments for the Western Consortium of Idaho, Nevada, and New Mexico”. In September 2008, NSHE (Nevada System of Higher Education) was awarded $15 million by NSF EPSCoR over five years to develop science, education, and outreach infrastructure at UNR, UNLV, DRI (Desert Research Institute), NSL (Nevada Seismological Laboratory), and NSHE community colleges for the study of climate change and its effects on Nevada. Then in September 2009 the Track II grant mentioned above was also funded by NSF EPSCoR to foster inter-state collaboration on related cyber-infrastructure developments.

The Nevada Climate Change project now comprises a web portal, the SENSOR data system (software, hardware, and database), and a high-speed TCP/IP network infrastructure. The NCCP (Nevada Climate Change Portal) website (http://sensor.nevada.edu) provides information to project members, researchers, and the public [2]. Via search interfaces and web services, users can search and download collections of SENSOR data. The web portal also provides real-time videos and photos from monitoring sites.

The SENSOR data collection system comprises numerous Campbell Scientific data loggers (CR1000 and CR3000), each with dozens of physical sensors that collect thousands of measurements per minute. Through the secure virtual private network (VPN) of the Nevada Seismological Laboratory [2], the data loggers transport collected data to the data center via the lossless TCP/IP protocol [3]. Then the data is stored on file servers and imported into a SQL server. By design, the SENSOR system uses a standard-neutral database schema, meaning that the database structure is not modeled after a specific metadata standard such as FGDC or ISO 19115 [3]. Furthermore, the system collects more information than is needed by any one standard. The main purpose of the work presented in this paper is to provide a system to transforms raw data to a chosen metadata standard.

As mentioned before, the NSF EPSCoR-funded project involves three Western states: Nevada, Idaho, and New Mexico. All these states have their own data centers or data repositories implemented using particular metadata standards. This creates the challenge: how can we effectively share data across data centers from all three states? One solution is to build a central clearinghouse to which each data center submits metadata for cataloguing [13]. Users can then search the clearinghouse for geographic data from all participating data centers. Each participant web portal can query the clearinghouse for data as well. This paper proposes a practical mechanism to transform files between different metadata standards. With this service, the clearinghouse will be able to handle data submissions accompanied by differing metadata standards, allowing each data center to utilize its own metadata standard, such as ISO 19115.

Specifically, in this paper, we propose to create a community metadata ISO 19115 adaptor that will transform geographic metadata in different standards to the ISO 19115 metadata standard format and vice versa. The project will be implemented using the SOAP web service technology, XML, XSLT, and C#.

3.2 Logical Model

The diagram in Figure 3 shows the logical model of the proposed approach. In this diagram, a user initiates a web service request with proper parameters. The web service will validate the request, record user information (optional), process the request, and return proper output metadata. The requested parameters could include, for example: geospatial information, sampling interval, time period, and return data format. The return data is an XML document that complies with the requested metadata standard. The web user interface shown in Figure 4 provides more details on how the user can interact with the proposed adaptor's functionality.

3.3 System Design/Software Workflow

Figure 2 presents the generic software workflow chart. The request can be a web form request or a script-generated web service request. The response is always an XML document in a metadata standard format.

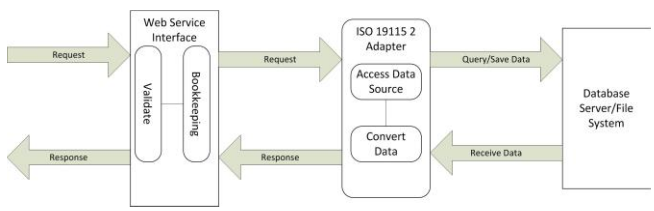

Figure 2: Software workflow with a web service interface module and an ISO 19115 adaptor module

3.4 User Interface

Users will utilize a form to easily construct the request. The web form will provide the necessary data for the web service and the ISO 19115 adaptor to properly query data and form (translate) metadata. For security purposes, a user must provide a valid user ID or developer reference number with each request. Although the web form user interface is easy to use, it is not the only way to access the service. A service request can be constructed by scripts that supply the same required request parameters. Figure 4 presents the user interface of the proposed adaptor.

3.5 Transformation Solution

In our proposed approach, the transformation between two different metadata standards is carried out using XSLT. That means two XSLT template files are required for each pair of source and target formats: $T_{a \rightarrow b}$ and $T_{b \rightarrow a}$.

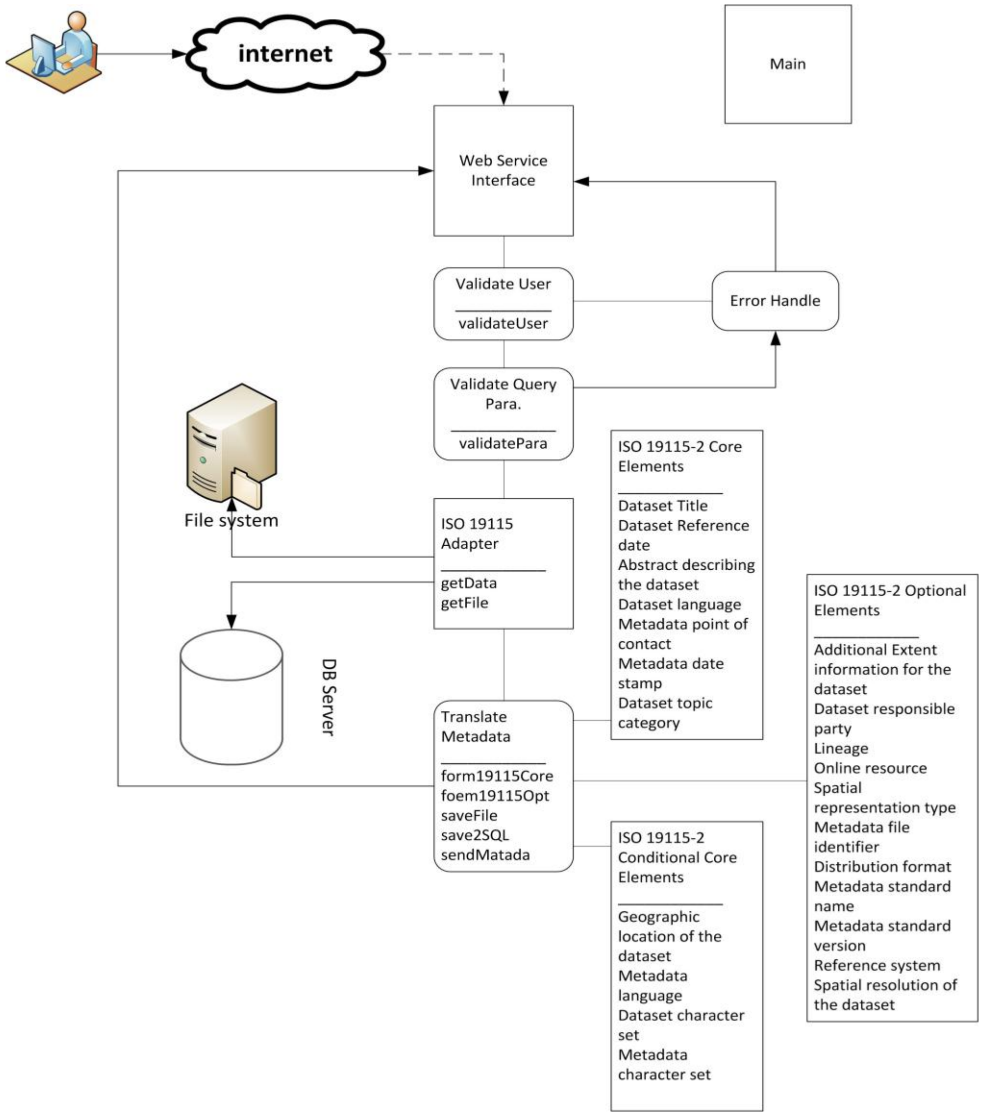

Figure 3: System design showing basic functionalities of the web service based ISO-19115 adaptor

This approach is straightforward when the number of supported formats (n) is small because the number of XSLTs needed to transform from a source format to a target format without using a transformation sequence is $T(n)=n!$

For example, 6 XSLTs are required for 3 supported formats. This mechanism works well with a small number of supported formats, but is clearly inefficient for a large number of formats. For example, $T(10) = 3,628,800$ and $T(15) = 1,307,674,368,000$. This is clearly impractical.

Some researchers have suggested using the graph theory. By letting each vertex in the graph represent a one-way transformation, as long as a connected graph is created, any source format can be theoretically transformed into any target format by using a transformation sequence. Furthermore, weights can be assigned to each edge based on the transformation complexity, computation time, etc. [1]. Then the graph theory (Dijkstra's algorithm) can be implemented to find the shortest path to improve efficiency.

Figure 4: Web user interface to construct and send a web service request to search for data or to request for a metadata standard transformation

However, the graph makes the concept complex. Given a set of transformations, determining the most efficient transformation sequence between source format A and target format B is essentially equivalent to solving the shortest path problem for a graph in which the data formats are represented by vertices, and the transformations are represented by edges [6, 8]. We propose a new mechanism: centralized transformation. We pick a popular format as the transformation center point (TC) and create the two-way XSLTs for each pair of formats in which one of the formats is the center point format TC. This way, only 2*(n - 1) XSLTs are needed. A transformation between any two formats $T_{n \to m}$ can be carried outwith a sequence of only two transformations: $T_{n \to m} = XSLT_{n \to TC} + XSLT_{TC \to m}$ (Figure 5).

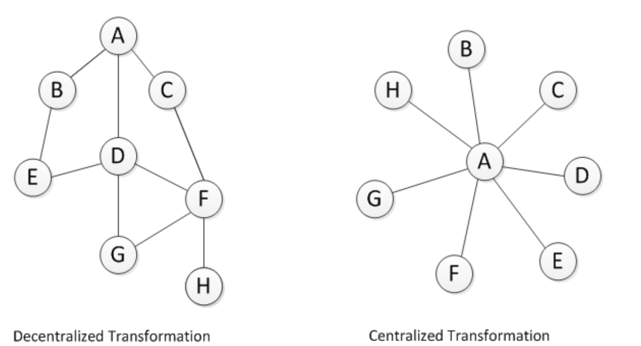

Figure 5: Comparison between graph-based transformation (on the left side) and centralized transformation (on the right side)

4 Related Work

4.1 GeoNetwork Opensource

GeoNetwork Opensource is a standards-based and decentralized spatial information management system designed to enable access to geo-referenced databases and cartographic products from a variety of data providers through descriptive metadata, enhancing the spatial information exchange and sharing between organizations and their audiences using the capabilities of the Internet [7]. The GeoNetwork Opensource is an open source Java application developed following the principles of Free and Open Source Software (FOSS) and based on the International and Open Standards for Services and Protocols. The project started out as a Spatial Data Catalogue System for the Food and Agriculture Organization of the United Nation.

4.2 IDACT Transformation Manager

The IDACT Transformation Manager project was part of a group research on the Arctic Pollution Issues conducted by AMAP (Arctic Monitoring and Assessment Programme). It leverages the knowledge of a group of users, ensuring that once a method for a data transformation has been defined (either through automated processes or by a single member of the group) it can be automatically applied to similar future datasets by all the members of the group. This prototype was primarily focused on the dataset itself instead of the metadata [6, 9].

5 Conclusions and Future Work

Data is valuable and expensive to collect. Organizations and institutes invest heavily in both equipment and labor to collect data, sometime just for a specific study. Thus, a method to reuse this valuable data becomes an important research topic. Different organizations use different software systems during the data collection phase. Therefore these data are in different formats, which make it hard to reuse, discover, and catalogue them. By introducing metadata standards to describe data, researchers can easily search and discover existing data for their studies. In the geographic domain, there exist multiple metadata standards. Thanks to the XML/XSLT technology, it is relatively easy and straightforward to represent those metadata standards in the XML format and perform different transformations among different metadata standards using XSLT.

In this paper, we proposed a mechanism for transformations between metadata standards based on web services and the XML/XSLT technology. This will enable more effective sharing of geographic data represented in different data and metadata standards. Future work will include implementing the full web service system and customizing it so it can be used in NCCP to generate ISO 19115 metadata on the fly.

Acknowledgments

This work was made possible through the support provided by the National Science Foundation under Cooperative Agreements No. EPS-0814372 and No. EPS-0919123.

References

[1] Wei, Y., Di, L., Zhao, B. et al. (2007). "Transformation of HDF-EOS metadata from the ECS model to ISO 19115-based XML." *Computers & Geosciences*, Vol. 33.2 , 2007, pp. 238-247.

[2] McMahon, M. J., Jr. et al. "Architecting Climate Change Data Infrastructure for Nevada." *Lecture Notes in Business Information Processing*, LNBIP-83, Springer, 2011.

[3] McMahon, M. J., Jr., Dascalu, S.M., Harris, F.C., Jr. and Strachan, S., "SENSOR - Applying Modern Software Data Management Practices to Climate Research", *Proceedings of the 2011 SNA Workshop*.

[4] Metadata standards, November 2012, Wikipedia, *http://en.wikipedia.org/wiki/Metadata standards*

[5] Metadata standards, 2012, DCC (Digital Curation Centre) *http://www.dcc.ac.uk/resources/briefing-papers/standards-watch-papers/what-are-metadata-standards*

[6] Hay, B. IDACT Transformation Manager. *Dissertation*, Montana State University Bozeman, 2005.

[7] Ticheler, J. & J. Hielkema. "GeoNetwork Opensource Internationally Standardized Distributed Spatial Information Management." *OS-Geo Journal*, 2.1 2007.

[8] Nance, K. L. and B. Hay. "Automatic transformations between geoscience standards using XML." *Computers & Geosciences* 31.9 (2005): 1165-1174.

[9] Nance, K. and B. Hay. "IDACT: Automating Data Discovery and Compilation." *Proceedings of the 2004 NASA Earth Science Technology Conference*.

[10] Greenberg, J. "Metadata extraction and harvesting." *Journal of Internet Cataloging* 6.4 (2004): 59-82.

[11] Kawtrakul, A., and C. Yingsaeree. "A Unified Framework for Automatic Metadata Extraction from Electronic Documents." *Procs. of the Intl. Advanced Digital Library Conference*. Nagoya, Japan, 2005.

[12] Nogueras-Iso, J. et al. "Metadata standard interoperability: application in the geographic information domain." *Computers, Environment and Urban Systems* 28.6 (2004): 611-634.

[13] Bigagli, L. et al. "GI-Cat: A Web Service for Dataset Cataloguing based on ISO 19115." *Proceedings of the 15th IEEE International Workshop on Database and Expert Systems Applications*, 2004.