Chapter 2

Background: Metadata exchange approach

We all heard the phrase "global warming". The Wikipedia article states that "Global warming is the rise in the average temperature of Earth's atmosphere and oceans since the late 19th century and its projected continuation. Since the early 20th century, Earth's mean surface temperature has increased by about 0.8 °C (1.4 °F), with about two-thirds of the increase occurring since 1980. Warming of the climate system is unequivocal, and scientists are more than 90% certain that it is primarily caused by increasing concentrations of greenhouse gases produced by human activities such as the burning of fossil fuels and deforestation. These findings are recognized by the national science academies of all major industrialized nations."[14]

As we can see in Figure 2.1, global warming is indeed occurring . The average temperature has been increasing from 1880 to 2012. The question is what caused global warming. Scientists are 90% sure that global warming is caused by human activities. There are scientists who do not agree. They believe it is normal to have a higher temperature during "summer" than "winter" of our solar system or even the universe.

On the other hand, decision-makers and policy-makers need more "knowledge" so they can establish rules (laws) to address the issue. Global warming is a world wide issue. It affects each and every one of us on the planet. The new rules will affect people's life and it costs money. So it must be preceded with cautions. Some questions need to be answered first. Is the increasing temperature caused by human activities? What activities? Is it reversible? Scientists can help developing their knowledge by building models, analyzing data, proving their models. This leads to the next element, data. Do we have the data? Do we have enough useful and reliable data? I think we do. A staggering amount of data, especially geographic data, is collected and stored everyday and the increasing rate at which data is being collected is overwhelming. In the meantime, scientists have developed complex models that can analyze and manipulate larger, richer, finely-grained datasets from multiple heterogenous sources than ever before. It provides an opportunity for researchers and scientists to investigate phenomena in great depth. One of the common problem researchers and scientists face is how to find and integrate heterogenous domain-specific data into a format that meets their own needs.

Researchers and scientists are highly trained professionals who can analyze data and transform data into knowledge for decision-makers. Based on the knowledge provided by scientists, decision-makers can make correct policies. That would mean that researchers and scientists should spend their valuable time and resources on analyzing data. But in reality, they need to spend a great deal of time on locating useful data, converting it from one format into the format to meet their needs while they could use the time more productively in analyzing those data.

Because companies and organizations use different equipment to collect data and use different software to store their data, researchers can not easily find the data they are looking for. A great amount of data remains untapped. Another scenario is that researchers invest heavily in equipment and software to collect data for a project. If there is not a system to make those data discoverable, it will be wasted while it could be reused by more data consumers.

Dataset itself is not useful whiteout the data about the dataset. It is like a backup tap is a piece of junk without a proper label. Metadata is used to describe data and provides a way to comprehensively describe datasets. It is widely used by data catalogues and clearinghouses. Various metadata standards have been developed by different organizations for different data categories.

Due to the coexistence of various metadata standards, making full use of metadata in different metadata standards becomes a problem.

This project addresses the metadata standards issue by providing a online web service to translated between different geographic metadata standards.

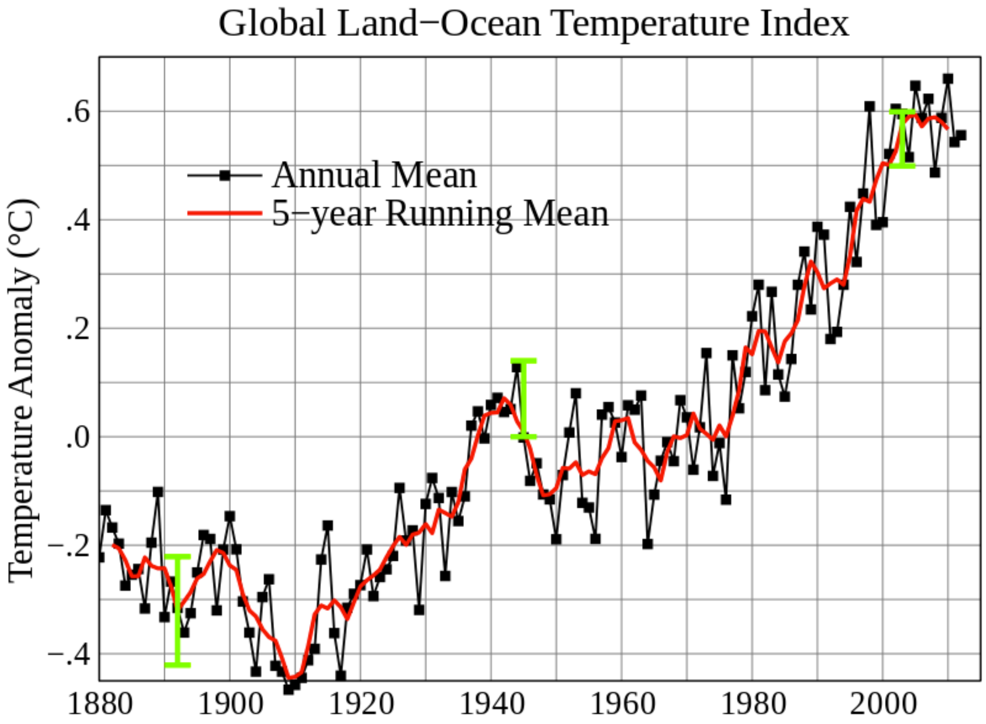

Figure 2.1: Global mean land-ocean temperature change from 1880–2012, relative to the 1951–1980 mean. The black line is the annual mean and the red line is the 5-year running mean. The green bars show uncertainty estimates. Source: NASA GISS

2.1 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 (transcript, text description of the scenes), and webpages (keywords, description, software used to create the page, author of the page).

The objective of metadata is to provide more details about a dataset. It has been widely used in dataset cataloguing and clearinghouse activities so data users can locate the data source more easily and accurately.

Metadata can be embedded inside the data file or stored in a separate document.

Furthermore, metadata is commonly used by computers and softwares, rather than humans. Metadata can be stored internally, in the same file as the data, or externally, in a separate file.

2.2 Metadata Standard

In the information age, the amount of data, especially geographic data, is exploding. Many companies produce and collect geodata in various formats from different sources with different equipment. The result of this heterogeneous geodata creates a major problem for data sharing. Geographic metadata describes the existing geodata. By reading the geo-metadata, users can get more information about the original dataset, like name, quality, ratio, data structure, etc. [11]. Because most geodata 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 geometadata 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 standards 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 15 classic metadata set, 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 defined elements, but it is hard to analyze and process. There is also not a single metadata definition language. Therefore, different standards are presented using different notations. For example, the ISO 19115 uses UML while the 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, 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].

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).

2.3 Geographical Metadata Standards

In the geographic domain, the most common metadata standards are FGDC's 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 19115's 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.

2.4 ISO 19115

The following is a definition of ISO 19115 from ISO website:
"ISO 19115:2003 defines the schema required for describing geographic information and services. It provides information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data.

ISO 19115:2003 is applicable to:

- the cataloguing of datasets, clearinghouse activities, and the full description of datasets;
- geographic datasets, dataset series, and individual geographic features and feature properties.

ISO 19115:2003 defines:

- mandatory and conditional metadata sections, metadata entities, and metadata elements;
- the minimum set of metadata required to serve the full range of metadata applications (data discovery, determining data fitness for use, data access, data transfer, and use of digital data);
- optional metadata elements - to allow for a more extensive standard description of geographic data, if required;
- a method for extending metadata to fit specialized needs.

Though ISO 19115:2003 is applicable to digital data, its principles can be extended to many other forms of geographic data such as maps, charts, and textual documents as well as non-geographic data." [15]

The ISO 19115 standard is part of the ISO 19000 series (ISO Geographic Information Suite of Standards). It defines how to describe geographic information and services, including contents, spatial-temporal purchases, data quality and right of use.

The objective of ISO 19115 is to provide a clear procedure to describe digital geographic datasets so that users can determine whether the data is useful and how to access it.

2.4.1 Tri-State Metadata Exchange

This effort is part of the metadata exchange project, which is, itself, part of Track II of the NSF EPSoR funded project "Collaborative Research: Cyberinfrastructure Developments for the Western Consortium of Idaho, Nevada, and New Mexico". On September 2008, NSHE (Nevada System of Higher Education) was awarded $15 million by NSF EPSoR 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. 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 collected from 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 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 standards-neutral database schema, meaning that the database structure is not modeled after a specific metadata standard like FGDC or ISO 19115 [3] – the system simply collected more information than was needed by any one standard. The main purpose of this work is to provide a system to transform raw data to a chosen metadata standard.

As mentioned before, the NSF EPSoR-funded project involves three western states: Nevada, Idaho, and New Mexico. Idaho and New Mexico have their own data centers or data repositories implemented particular metadata standards. This creates a problem: how can we effectively share data between data centers? 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 too.

This project proposes a practical mechanism to transform 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.

In this project, I 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 standard SOAP web service technology, XML, XSLT, and C#.