4.10 AUTOMATED SCANNING OF OBSERVATIONAL DATASETS FOR THE GENERATION OF FORMAL METADATA

Jason E. Fabritz and Donald W. Denbo*
University of Washington/JISAO, NOAA/PMEL, Seattle Washington

1. INTRODUCTION

Web-based data discovery presently uses search engines that index FGDC metadata. Metadata needs to be of high quality (few errors) and rich (contain many keywords and other descriptors) for a search to find the records of interest (and not too many extraneous records). High quality and rich metadata can be incredibly time intensive to produce by hand from observational data sets that can have thousands to millions of individual observations. In addition, observational data sets often lack the information necessary in the data files to produce rich metadata. Additional metadata needs to be manually added by the researcher to produce quality metadata. The Climate Data Portal (Soreide et al., 2002; Soreide et al., 2001), a distributed data system with data discovery capabilities, is one such system that can benefit from automated metadata generation.

We have constructed the foundation for a framework of tools that can scan observational data to produce formal FGDC metadata. The framework provides a utility that enables the user to open and edit a configuration file that describes collections of observational data files plus additional metadata. Next, a set of utilities process this configuration file, examine the associated observational data files, and insert the extracted metadata into a central metadata database. Other framework processes then extract information from this central metadata database to create metadata adhering to the FGDC Content Standard for Digital Geospatial Metadata (CSDGM) or other formats as required.

2. DESIGN GOALS AND DECISIONS

There are many potential ways to construct an automated framework for generating metadata. The principal design goals guiding the selection of technology and design were:

Flexibility, Robustness and Scalability. The framework must be flexible enough to gracefully adapt to changes in components and operating requirements. Also, It is desirable to produce a system that will be able to grow and fulfill future needs that are not visible at this time; including additional requirements such as ingesting additional observational file formats and producing metadata in multiple formats. One way to help isolate various components of the framework from changes of this nature is to use XML and Extensible Stylesheet Language Transformations (XSLT) for translating information from one format to another reducing the amount of refractoring necessary to adapt framework components to changes.

Portability. The computing environment at PMEL is heterogeneous in nature. It is necessary to produce a framework that can be deployed on a range of different hardware and operating systems. Java is by far the best-suited programming environment meeting this requirement.

Cost. It is desirable to use technology that leverages public or existing software licenses, technology and developer experience. MySQL is a robust and free database for non-commercial use and was already in wide use in the laboratory.

3. ARCHITECTURE

Flow of data through this processing framework can be divided into two pathways: automated metadata collection and upload, and metadata extraction with generation of formal metadata files. The collection and upload of new metadata takes place in four steps. First the software reads a configuration eXtensible Markup Language (XML) file (created by the operator using an interactive user interface, see Figure 1) that describes the global metadata properties and locations of observational data files to analyze.

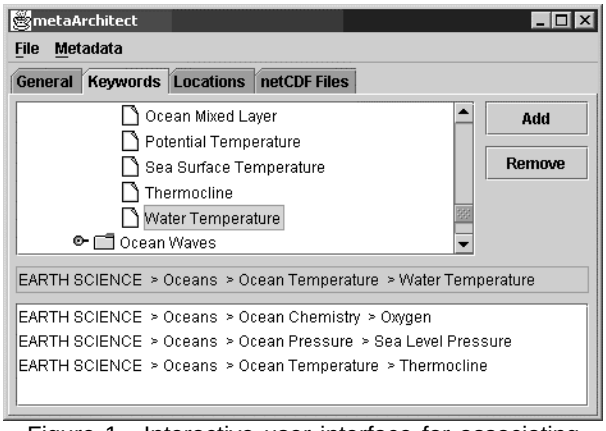


Figure 1. Interactive user interface for associating external metadata information with data files. Certain information such as contact information, terms of use and some keywords are not contained within the observational data and span over

* *Corresponding authors addresses:* Jason E. Fabritz, c/o JISAO, Box 354235, University of Washington, Seattle, WA 98195; e-mail: jfabritz@u.washington.edu. Donald W. Denbo, NOAA/PMEL/OCRD, 7600 Sand Point Way NE, Seattle, WA 98115; e-mail: dwd@pmel.noaa.gov.

multiple datasets. Therefore this information must come from an external source.

Next, the required observational data files (netCDF) are loaded and examined by automated routines for additional metadata information. This information is then combined with the specified global metadata and the results are written to a single observational-centric metadata XML file.

Next the framework applies an Extensible Stylesheet Language Transform (XSLT) converting the observational-centric metadata XML file into a metadata-centric XML file that closely matches the structure of the central metadata database store. Finally, an additional process reads the metadata-centric XML file, merging the new information into the central metadata database.

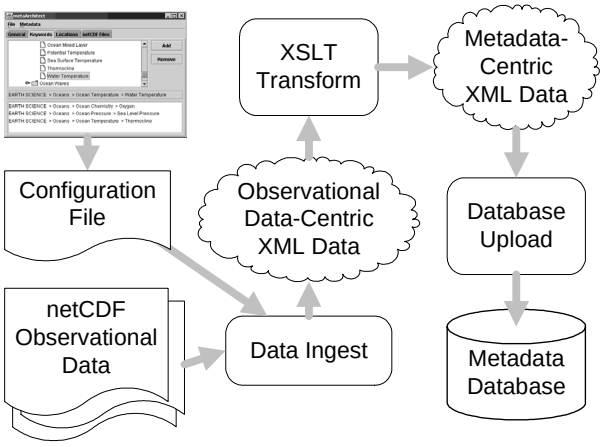
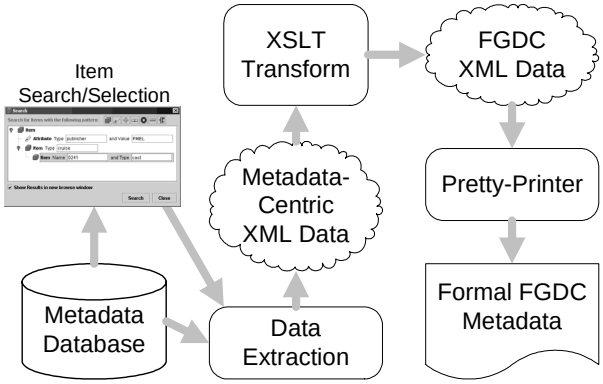


Figure 2. Illustration of the metadata ingest/upload process.

The generation of formal metadata takes place in five major steps. First, the unique identification number of the desired metadata item is obtained either by use of a browsing utility or specific direct query into the metadata database. Next a routine is called that interfaces with the central metadata database and extracts the requested metadata information for the item in question plus information about its children items and parent items. This information is output in the form of XML formatted specifically to be item-metadata-specific; the format of which is very similar to the metadata-centric XML format used in the upload process. The item-metadata-centric XML file is then converted via an XSLT transform into a FGDC formatted XML file (which is a different format than the FGDC Metadata DTD 3.0.1 19990611 format authored by Peter N. Schweitzer (U.S. Geological Survey, Reston, VA 20192)). Next, the FGDC formatted XML file is read by a routine that converts the XML file into a properly formatted FGDC file. As an extra precaution, the resulting FGDC file passes through a final routine that is a Java port of a subset of the MP compiler that verifies the structure of a formal metadata file.



Figue 3. Illustration of the formal metadata generation process.

4. FUTURE DIRECTIONS

As this framework matures, we plan to add more sophistication and richness to the resulting metadata product. This includes increasing the detail of the description of spatial extent of the data files and expanding the range of observational dataset formats that can be read and analyzed by the framework. Improvements to operator tools providing graphical user interfaces for administering the system are also in progress.

5. ACKNOWLEDGMENT

This publication is partially funded by the Joint Institute for the Study of the Atmosphere and Ocean (JISAO) under NOAA Cooperative Agreement No. NA17RJ1232, Contribution # 871. PMEL contribution 2407. The views expressed herein are those of the authors and do not necessarily reflect the views of NOAA or any of its sub agencies.

6. REFERENCES

Soreide, N.N., C. Sun, B.J. Kilonsky, D.W. Denbo, W.H. Zhu, and J.R. Osborne, 2002. A Climate Data Portal. In *18th International Conference on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*, Orlando, FL, 13–17 January 2002 (submitted).

Soreide, N.N., C. Sun, B.J. Kilonsky, D.W. Denbo and W. Zhu, 2001. A Climate Data Portal. In *17th International Conference on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*, Albuquerque, NM, 15-18 January 2001, 191-193.