The LBA-ECO Metadata Warehouse and Its Implications for Data Mining Initiatives

Lisa Wilcox, lwilcox@pop900.gsfc.nasa.gov (Presenter)
Amy L. Morrell, amorrell@pop900.gsfc.nasa.gov
Peter C. Griffith, peter.griffith@gsfc.nasa.gov

Organization: Science Systems & Applications, Inc. and the LBA-ECO Project Office, NASA Goddard Space Flight Center

Objectives of the System

In 2005, the LBA-ECO Project Office developed a system to harvest and warehouse metadata resulting from the Large-Scale Biosphere Atmosphere Experiment in Amazonia[1]. The harvested metadata is used to create dynamically generated reports, available from our website, that facilitate access to LBA-ECO datasets. The reports are generated for specific controlled vocabulary terms (such as an investigation team or a geospatial region), and are cross-linked with one another via these controlled vocabulary terms. This approach creates a rich contextual framework enabling researchers to find datasets relevant to their research. It maximizes data discovery by association and provides a greater understanding of the context surrounding each dataset. It also creates numerous starting points for a web crawler to harvest data for a data mining project.

For example, our website provides profiles for each LBA-ECO investigation. Each profile describes investigation participants, abstract(s), study sites, and publications. Also linked from each profile is a list of associated registered dataset titles and where applicable, publications that have used those datasets. Each one of these dataset titles in turn link to a dataset profile, which describes the associated metadata in an easily readable format. All of the reports mentioned here are dynamically generated from information stored in a database. The dataset profiles are generated from the harvested metadata, and are cross-linked with associated reports via controlled vocabulary terms such as those described above.

One example of an associated dynamically generated report is a profile for a region. The region name appears on the dataset profile as a hyperlinked controlled vocabulary term. When researchers click on this link, they find a list of reports relevant to that region, including a list of dataset titles associated with that region, organized by site and investigation. Each dataset title in this list is hyperlinked to its corresponding dataset profile. Moreover, each dataset profile contains hyperlinks to each associated data file at its home data repository. A researcher interested in mining data pertaining to a particular LBA-ECO region could start by configuring a web crawler to harvest the data linked from these reports.

We plan to add to our metadata warehouse by harvesting metadata associated with the North American Carbon Program (NACP)[2]. This is particularly challenging because

NACP is an interagency endeavor drawing on datasets in many different formats, residing in many thematic data centers and also distributed among hundreds of investigators[3]. These challenges will certainly enhance our expertise in processing metadata from disparate datasets. Additionally, scientists and program managers will benefit because an NACP metadata harvest with access to interdisciplinary data is required to complete the creation of a North American carbon budget.

We would also like to add search and retrieval capability allowing researchers to search metadata within both projects. The returned results would link to a record that is similar to the dataset profiles and would contain links to the data files housed by investigators or in thematic data centers. It is hoped that this application will provide centralized access to the data from both projects. This has implications for data mining in that our researchers could work with us to create a customized report listing datasets of interest from both projects. A web crawler could then be configured to harvest the data linked from this report. This is another potential starting point for a data mining project.

Technical Overview

Metadata for LBA-ECO datasets reside at two repositories: the Centro de Previsão de Tempo e Estudos Climáticos (CPTEC) in Brazil and the Oak Ridge National Laboratory (ORNL) Distributed Active Archive Center (DAAC) in the United States. Each of these metadata records is available online as an XML file. A list containing a file location for each record is provided to the system by a separate process. The harvesting process inputs this list to the Swish-e[4] spider and retrieves the files into a single ASCII output file.

After the harvesting process is complete, our parsing script ingests the contents of the output file into the MySQL[5] database developed for the metadata warehouse. The database was designed to be compliant with the Content Standard for Digital Geospatial Metadata (CSDGM), a widely accepted metadata standard for geospatial data[6]. The parsing script is based on a crosswalk that maps the harvested metadata to this metadata standard. It is written in PERL, and makes extensive use of the XML::Simple CPAN module[7] to parse the metadata. All of the scripts that create our dynamically generated reports are also written in PERL, using PERL's DBI interface to access our MySQL database.

Implications for Data Mining

We are "mining" metadata to facilitate scientists' online access to data, as opposed to mining data to answer science questions. Our work has important implications for data mining initiatives in that the applications and reports developed from the metadata warehouse facilitate online access to data in various ways. This is an essential first step to the data mining harvesting process. In essence, it is a web portal to the "mine". We can also work with our researchers to create customized, harvestable reports tailored to their research needs, as well as assist them in creating their own harvesting processes for data mining. Similar projects can use these tools and techniques to

develop their own web portals. They can also use these ideas in partnering with an existing metadata clearinghouse.

---

[1] Large-Scale Biosphere Atmosphere Experiment in Amazonia (LBA-ECO), http://www.lbaeco.org.
[2] North American Carbon Program (NACP), http://www.nacarbon.org.
[3] *Data Management for the North American Carbon Program -- Workshop Report*, New Orleans, LA, January 25-27, 2005, http://www.nacarbon.org/nacp/documents/NACP_DataMgmt_new.pdf.
[4] Simple Web Indexing System for Humans – Enhanced (Swish-e), http://swish-e.org/.
[5] MySQL, http://www.mysql.com.
[6] Content Standard for Digital Geospatial Metadata (CSDGM), http://www.fgdc.gov/metadata/contstan.html.
[7] Comprehensive Perl Archive Network, http://www.cpan.org/.