# Content Based Metadata Workbench (CBMW)

**Ramachandran Suresh (Mayur Technologies, Silver Spring, MD 20905)**
**Joel Sachs (UMBC, Baltimore, MD)**
**Robin Pfister, Jeanne Behnke (GSFC, Code 586, Greenbelt, MD 20771)**

## Introduction

Currently NASA's Earth science data systems treat information as datasets, storing them off-line and only making file metadata available for search and retrieval. As a result, useful data content (e.g. geophysical parameter values) is hidden from end users during the data access process. Access systems are limited to queries on the existence of geophysical parameters. E.g. "show me all data collected over the Atlantic Ocean that contain sea surface temperature data". We cannot currently support content-based searches such as "show me all primary productivity data collected over the Atlantic Ocean where sea surface temperature is greater than 25 C". Additionally, scientists spend significant resources (time and money) ordering and analyzing data just to identify features of interest to form the basis for scientific research.

To enable better access to and use of Earth Science information, our team is prototyping the Content Based Metadata Workbench (CBMW) [1]. CBMW seeks to improve the utility of information hidden in NASA's Earth science data systems by providing a global geophysical parameter data warehouse that is interoperable with metadata search systems. It will provide the following: a mechanism for content-based metadata searching of Earth science data, a research planning tool that includes inter-parameter visualization, and a broad warehouse of meaningful Earth science data to serve as a target for data mining.

## Background

The goal of the CBMW is to improve the utility of information hidden in NASA's Earth science data systems by providing a global geophysical parameter data warehouse that is interoperable with metadata search systems to serve three primary objectives. The first is to provide a mechanism for content-based metadata searching of Earth science data. The second is to provide a research-planning tool that includes inter-parameter comparison. The third is to investigate a broad repository of meaningful Earth science data to serve as a target for data mining.

## Operations Concept

Figure 3 illustrates the operations concept of the CBMW. For research planning, users can display geophysical parameters of interest for specific locations and time. A variety of visualization mechanisms are envisioned so that the users will have many ways to view the parameters and identify cross-parameter characteristics such as interesting anomalies, for further research. To extend this scenario to content-based searching, the user can select features of interest from the display, then select criteria for other metadata attributes describing datasets available at the archives, and send a search to the archives for other data that is coincident with the selected feature. Coincidence is established on the space and time footprint associated with the pixels that were used in the data display where the user selection was made. Although not a primary goal of this effort, because these parameters are available on-line, they can be available to data mining tools for mining exercises.

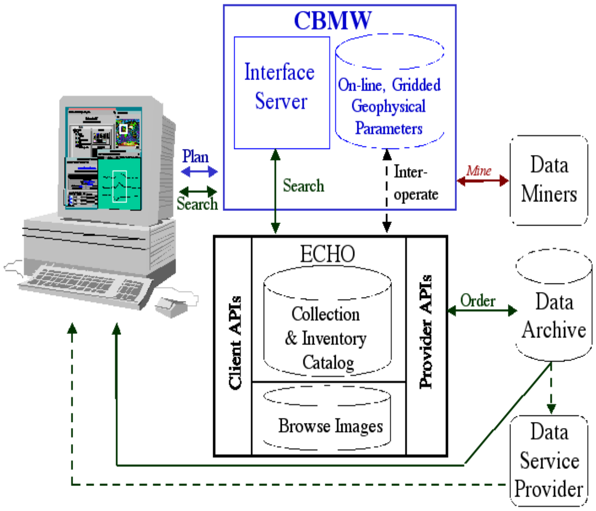

**Figure 3.** Content Based Metadata Workbench Concept and Context.

## Approach

Level 3 and 4 products held in NASA's archives contain geophysical parameter values that have been derived based on established and broadly accepted algorithms. Since these products are relatively small, our intuition was that today's technology would make it feasible to put them on-line to enable advanced search techniques and support interactive application of functions and routines. This turned out to be true, to an extent. A level 3 product typically contains one or more geophysical parameters, in addition to several ancillary parameters, such as solar azimuth angle. To keep the CBMW small, we chose not to store ancillary data. High-level products are passed from external archives to the CBMW through various mechanisms (preferably via subscription and FTP).

We considered three approaches for building CBMW: one based on RDBMS technology; one based on markup languages; and one based on OGC tools. Because of the diversity of MODIS data products, and also because MODIS level-3s are in HDF-EOS format, our primary focus was on products derived from MODIS measurements. MODIS [2] products are logically categorized as land, ocean, or atmosphere. There are some differences between these three types of MODIS level 3. In the land domain, one dataset typically presents one geophysical parameter, derived by an individual PI and his/her science team. In the ocean domain, there exist several "joint" products, where several geophysical parameters are stored as separate scientific datasets in the same HDF-EOS file. In the atmosphere domain, there is a single joint data product, comprising over 650 parameters. About 100 of these are geophysical parameters, with the rest constituting ancillary data. (Characterizing a datum as ancillary depends, to an extent, on point of view. Cloud cover (a geophysical parameter in its own right), for example, is ancillary to a vegetation index dataset.) The size of the atmospheric (800 MB) product has been a barrier to the effective use of the information that it contains.

Another difference between the data products of the three domains is their choice of map projection. Land data is in the integerized sinusoidal (ISIN) projection, ocean data is in geographic, and atmospheric data can be either geographic, or equidistant cylindrical. The user of CBMW should not have to worry about this. Typically, she will want to use latitude and longitude as coordinate axes, and so data should be represented to her in a geographic or equidistant cylindrical projection.

RDBMS Approach

The data files containing pixel-level geophysical parameters are read into the CBMW such that each pixel occupies a record in the database. Additionally, metadata on specifics of the granule, including the location, date and time of each pixel are stored in the database. Having stored multiple geophysical parameters in the database, we can rely on the expressive power of SQL to issue a wide range of content-based queries.

Searching content across products of varying spatial and temporal resolution require reprocessing of the data onto a common-space time grid, since, in order for the information displayed to be meaningful, these must be compared on a common basis. For the CBMW, this process begins in dataset selection, where we focus on 1 degree x 1 degree datasets.

To ingest heterogeneous datasets into the database, we rely on reprojection and transformation. To deal with temporal scale disparities, we record the start and stop time of each granule in the database, and rely on MySQL's ability to do timestamp comparisons. For example, if we ask for data for June 15, the database is queried for granules that begin before June 15, and end after June 15. For consistency and efficiency, we use weekly datasets where possible.

The CBMW will be interoperable with archives via the EOS Clearinghouse (ECHO) system. Specifically, CBMW returns a list of bounding boxes that satisfy a user's query (e.g. "where is SST is greater than 25 deg?"). CBMW will also transform this list into an ECHO request, according to the user's query (e.g. get me AVHRR level 2s where SST is greater than 25 deg.) In other words, to fully answer the user's query, CBMW becomes an ECHO client. ECHO's metadata clearinghouse will serve as the data search mechanism. ECHO will also broker requested services (e.g. subsetting, processing) as specified by the user if the user wants services applied before ordered data are delivered.

Markup Languages.

A debate currently rages in the database research community. The point at issue is how to deal with XML – through building native XML stores [3], or through accommodation of XML data structures in an object-relational setting [4, 5]. However there is still no accepted way to use markup languages effectively for content-based search since so much of the data is in binary format [6, 7]. This may be possible if we write an algorithm to identify geophysical parameters and anomalies using a Markup language. Pursuing this approach was beyond the scope of our prototype project. However, dealing elegantly with binary data will bean essential step to the realization of a semantic web for scientific data, and so we hope that others will shortly be taking up this problem.

OGC

At this time, there are no software tools to support searching earth science data using OGC standards. However, we studied several commercial and academic projects that are engaged in providing better search capabilities. It is appropriate that the CBMW be developed in a manner that takes advantage of OGC standards to enable interoperability in the future. It is appealing to rely on a GIS to provide the overlay and intercomparison features that we desire.

## Challenges

There are several technical challenges associated with building the CBMW. Some of these we overcame in construction of the prototype, others remain to be addressed. One is how to best store the content information on-line so that it can be readily searched and visualized. Another is in understanding the problems associated with storing and searching content-based data. This includes a real focus on the data management issues. Another challenge is interoperating with the other archives so that search criteria defined based on CBMW content can return non CBMW products.

Many technical challenges come about because of the nature of the data products. For example, some data granules are enormous, making ingest into the CBMW impossible without some sort of preprocessing of the granule. Specifically, we reproject to a 1 degree x 1 degree geographic projection, and we ingest only a subset of the parameters present in a dataset. In general, we do not ingest ancillary data (e.g. solar azimuth angle), and instead extract only targeted geophysical parameters. Data fusion presents another challenge. Searching content across products of varying spatial and temporal resolution and parameters with varying units of measurement will be a challenge because in order for the information displayed to be meaningful these must be compared on a common basis. Related to this issue is the visualization of diverse information in a user interface to support a navigation and discovery paradigm for data access. Currently, users are expected to manually interact with a website during search. Enabling this automation is the point of the impending semantic web.

One of the biggest problems CBMW faces is the effort to ingest Level 3 products in a uniform manner into a common database schema. No single ingest program can be developed to do this task. It represents the most time-consuming step in the CBMW effort. Projections represent another major challenge to developing a content-based metadata workbench. The data are in such differing projection that CBMW needs to determine how best to support searches across multiple projections.

## Summary

Content-based search of data is an issue that needs to be addressed now by science data management teams. Users have outgrown current search capabilities and are looking for faster, more efficient ways to access data in large archives. The CBMW is a continuation of the effort to provide better access to data. In the first phase, the CBMW has looked at several ways to design content-based search systems. This effort has shown that there are many issues that need to be considered as part of a development effort. It has also concluded that it is feasible to build a CBMW on relational database technology, even though this may not be the optimal choice. The results of the phase I effort are documented in a formal report which is available from the authors. There is a critical need for a mechanism for content-based metadata searching of Earth science data, a research planning tool that includes inter-parameter visualization, and a broad warehouse of meaningful Earth science data to serve as a target for data mining.

## References

1. The Content Based Metadata Warehouse http://samogon.gsfc.nasa.gov:7070/cbm_indexnew.htm

2. MODIS Science Data Support Team http://ltpwww.gsfc.nasa.gov/MODIS/SDST/

**3. The Timber Native XML Database http://www.eecs.umich.edu/db/timber/**

**4. XPERANTO: Publishing Object-Relational Data as XML http://citeseer.nj.nec.com/carey00xperanto.html**

5. Bridging Relational Technology and XML http://db.cs.berkley.edu/dblunch/jayavel.ppt

6. Handling Binary Data in XML Documents http://www.xml.com/pub/a/98/07/binary/binary.html

7. Discussion on Binary XML Proposalshttp://lists.xml.org/archives/xml-dev/200104/threads.html#00205