# The Role of Metadata Standards in EOSDIS Search and Retrieval Applications

Robin Pfister
NASA/Goddard Space Flight Center
Code 423
Greenbelt MD 20771
Ph: (301) 614-5171/ Fax: (301)614-5257/ e-mail: robin.pfister@gsfc.nasa.gov

## INTRODUCTION

Metadata standards play a critical role in data search and retrieval systems. Metadata tie software to data so the data can be processed, stored, searched, retrieved and distributed. Without metadata these actions are not possible. The process of populating metadata to describe science data is an important service to the end user community so that a user who is unfamiliar with the data, can easily find and learn about a particular dataset before an order decision is made. Once a good set of standards are in place, the accuracy with which data search can be performed depends on the degree to which metadata standards are adhered during product definition. NASA's Earth Observing System Data and Information System (EOSDIS) provides examples of how metadata standards are used in data search and retrieval.

## DEFINITIONS

There are generally two levels of metadata specificity required to describe science data. Collection-level metadata describe whole datasets, or collections of data items (e.g. Landsat Thematic Mapper Dataset). Granule-level metadata describe the smallest unit of data that is independently managed in the inventory (e.g. a scene of Landsat Thematic Mapper data). A granule, or data item, usually is a single profile, or scene of data but data granularity can vary depending on the nature of the science data, the purpose for which it was collected, and data organization preferences of the data producer.

There are four basic categories of metadata used in search and retrieval systems – Directory, Inventory, Guide and Browse. Directory metadata provide brief descriptive information about a whole dataset so a user can decide whether or not a dataset is of particular interest to pursue for potential order. Directory contains only a subset of the collection-level attribute metadata for a dataset as well as a brief description which is more along the lines of a short document. Inventory metadata are the core driver of the dataset search and retrieval system in that they allow users to identify specific data items for potential ordering. Inventory metadata include all collection-level attributes plus granule-level attributes that describe individual data items. Guide metadata provide users detailed descriptive information about datasets, platforms, instruments, campaigns and data centers and can provide links to other supporting documentation. Browse metadata provide a visual representation of a sample data granule that indicates the nature of the content. In most cases, browse metadata represent the specific data granule in question, but it may be only a sample that has spatial or temporal characteristics different from the data granule in question. Browse metadata can be in a variety of forms such as images, histograms, and contour plots.

## ROLE OF STANDARDS

The roles of standards in each of the four categories of metadata vary due to the different purposes and access methods in a search and retrieval system. Attribute/value metadata that reside in databases require consistency and accuracy in naming conventions. This ensures that the user gets all of the data that meet the query criteria and only the data that meet the query criteria - results are as complete and concise as possible. Metadata in the form of documents need to have a consistent layout so that users can generally find similar information in consistent locations of the document.

In general, directory, inventory and guide metadata are searchable by keywords or text strings. Inventory and some directory metadata contain attribute/value database fields. Guide and some directory metadata take the form of documents and are directly searchable by keywords and text strings in addition to being available through links once a dataset is identified. Browse metadata are usually retrieved through a link or access request after a specific granule of data is identified. In other words, the user first searches the inventory, then accesses the associated browse metadata through a link. Some systems use image type browse metadata as the primary discover mechanism to identify data items for retrieval. However, in order to get adequate user interaction performance, the system must hold or pull the browse metadata to a site that is local to the user before viewing. This process may be cumbersome.

### Directory Metadata Standards

Directory metadata typically contain both attribute and document components. Thus the standards are driven by the need to provide consistent and accurate results in keyword-based searching and by the need to provide users a consistent document layout. The directory standard used internationally by most organizations is the Directory Interchange Format

(DIF). More information can be found on the Global Change Master Directory (GCMD) web site at http://gcmd.gsfc.nasa.gov.

### Inventory Metadata Standards

Inventory metadata include collection- and granule-level metadata and can contain core and product-specific attributes and values. Core metadata are attributes that are generally common across most datasets in the database (e.g. all datasets have a granule identifier which is a unique key for identifying each data granule within a particular dataset). Product-specific metadata are attributes that are unique to a particular dataset. Inventory metadata standards include attributes and the format for their content as well as valid values (referred to as "valids") or valid ranges where applicable. They should also include naming standards for valid values as well as for product-specific attributes. It is essential that definitions are available for all attributes and valid values so that end users can be clear about the meaning of terminology. EOSDIS has a comprehensive suite of inventory metadata standards. Information on EOSDIS metadata standards is available at http://ecsinfo.hitc.com/applied_tech/metadata.

### Guide Metadata Standards

Superficially, guide standards focus on consistent content and layout of guide documents, however the standards that support effective search and retrieval vary depending on how documents are linked to data. Preferably, documents are linked to data through reference locators held in the database management system (DBMS) with the inventory data. However, in some cases such as EOSDIS, the guide system was developed independently of the archive inventory so such links do not exist. When database links do not exist, it is important that valid values and ranges for pertinent collection-level attributes are included inside the document so that keyword searches result in reliable results. Often the text where these are specified can be hidden as to not burden the user who needs to read or print the document. The EOSDIS Version 0 prototype included an independently developed guide system that identifies five guide document types. These describe datasets, platforms, instruments, campaigns, and data centers. Standards for each can be found at http://harp.gsfc.nasa.gov/v0ims/guideshells.html.

### Browse Metadata Standards

Like document metadata, browse metadata generally are linked to associated data through a reference locator held in the inventory database. Browse metadata exist in a variety of formats depending on the nature of the system in which these metadata are to be made accessible. The EOSDIS core system specification for browse metadata can be found in the "Browse Granule Description" document which can be located by performing a "quick search" on the phrase "browse" at the EOSDIS documentation web site (http://edhs1.gsfc.nasa.gov/). In EOSDIS, the browse standard is HDF-EOS. More information on this standard is available at http://hdfeos.gsfc.nasa.gov/hdfeos/workshop.html.

Browse metadata are most useful when they can be made available in an interactive search and access session. However some browse metadata are only accessible via FTP. This requires the user to go outside their search session to browse sample images before they order the data. In order to be useful to users, browse metadata need to include browse descriptions (how the browse image was produced), legends (display color/pattern to data value mapping), and location information.

## METADATA IN SEARCHING

### Distributed versus Central Metadata Repositories

In architecting a distributed or federated search and retrieval system, a decision must be made as to where the metadata inventory should be held – in distributed repositories near the data, or in central repositories potentially away from the data. With the development of standards such as XML and with improved networks this may become less and less of an issue but in today's environment the kinds of search services provided by the system still depend on the proximity of metadata to the user interface. Some systems are developed to give a perception of being fully distributed when in fact they are not. For example, web search engines (e.g. Lycos, Yahoo etc.) appear to be fully distributed, although they are actually built on a centralized repository of some form of metadata that has a reference (URL) to metadata stored at remote sites.

There are two basic ways for users to find data in information systems. In one, the user specifies query criteria and issues a search to the system. The system responds to the query with a list of data that meet the users criteria. The user may hone the results list but ultimately orders from that list. In this paradigm, distributed and central repositories are equally effective. The EOS Data Gateway (EDG) system (http://eos.nasa.gov/imswelcome) provides an example of this. The second paradigm is where the metadata are displayed to the user by the system and the user simply goes down a path by making sequential selections that ultimately lead to the actual data. In this case, the only request that goes to the archive is for access to the specific data item(s) that the user wants to retrieve. There may be local queries running behind the scenes, but the user is unaware of this interaction because the system is formulating the queries on behalf of the user. In this latter case, the chance of zero-hit results are reduced to zero. However, for this paradigm to be effective in terms of performance, metadata should be held at, or retrieved to, a single site. The Human-Computer Interaction Laboratory at the University of Maryland in College Park, Maryland has explored visualization of this paradigm in a prototype that can be found on their web site at http://www.cs.umd.edu/hcil/eosdis/index.html. An implementation of this second case can be found on the GCMD web site at http://gcmd.gsfc.nasa.gov/. Individuals tend to prefer one interaction style over the other but the preference is nearly equally divided between the two styles. So system architects should keep this in mind during information and systems architecting.

Some special search functions require that metadata be close at hand. One example is coincidence searching. Coincidence searching is finding data from different datasets that cover the same region at the same time (with some degree of tolerance). This type of searching requires a high degree of automated iteration of metadata comparison and therefore requires that the metadata be in a single location. Content-based metadata search and research planning tools are other functions that benefit from locally co-located metadata.

### Criteria Specification

Searches against attribute and valid value metadata from DBMSs yield more accurate results than keyword or text string searches of indexed documents. This is why it is preferred that documents are linked to the data through a reference locator (e.g. URL) held in the inventory database. In addition, the specification and availability of valid values for attributes during search construction can help prevent the formation of non-sensical queries by the user and can reduce zero-hit results. This can be done through "dependent valids". "Dependent valids" is the system's use of known relationships between valids and their associated datasets to help narrow the possible selections for the user. So once the user selects from a list of valid values for a particular attribute, when the user goes to specify criteria for the next attribute, the list is narrowed based on the applicability to datasets associated with the previously selected criteria. Dependent valids also allows for smart routing of queries because relationships between datasets and archives are known. Searches go only to archives that hold the datasets for which valids are specified.

Since product specific attributes apply to individual datasets, it is important that these are treated the same as dependent valids. Once a product-specific attribute is specified in the criteria, only other product specific attributes from that dataset should be displayed for selection.

### Metadata Exceptions in Querying

Core metadata attributes are common across many datasets, so specifying core criteria is usually straightforward. However, it is possible that an attribute doesn't exist for a particular dataset. In addition, it is possible that within a dataset, for a particular attribute, the values generally exist but in some cases there may be unpopulated (null) values. This introduces a degree of ambiguity into the metadata and thus into the query. To minimize this situation, it is important for data producers to populate all applicable attributes to the fullest extent possible. However, there may be some cases where these attributes simply do not apply, or the values are simply not available. In these cases query systems can offer the user the option to include or not include data where these metadata are unspecified.

### Content-Based Metadata

Research planning tools that allow users to plot and analyze content-based metadata (e.g. climatology, cloud content, etc) to identify specific data features by which to order data require that content-based metadata be collected and stored in the inventory. Content-based metadata can range from simple single attributes which give an overall indication of the content of the granule to more complex metadata such has histograms, tables or indexed high-level processed data products. Currently, few data producers generate content-based metadata but there are some efforts underway to construct data warehouses for content-based metadata.

### Results Honing

Users hone query results through metadata analysis. Tools that facilitate this process are extremely valuable. They range from simple tools such as metadata sorting utilities to the more complex research planning aids. Regardless, the usefulness of these tools depend on the degree and quality of metadata population. Another extremely useful function is incremental search. This gives the user the ability to specify new search criteria to apply to a results set rather than going back out to the archives for querying.

### Screen Configurability

It quickly becomes obvious that in a system with hundreds of core metadata attributes and thousands of product specific attributes to choose from, it is important to allow the users to configure the contents of the search and results screens. It is useful from a user perspective in that it allows the user to focus on the information that (s)he finds important. It is useful from a system perspective so the search engine does not waste resources searching for unnecessary metadata.

## SUMMARY

Successful data search and retrieval as well as effective metadata analysis depend on well populated, accurate and consistent metadata. End users depend on the extent, quality and content of metadata in order to find, access and use data products. This can be achieved through the specification and use of good metadata standards.