SH51A-0261

Design Considerations for Data Catalogs

J.A. Hourclé (RSIS/NASA-GSFC)

A.R. Davey (SAO)

I. Suárez-Solá (NSO)

K.Q. Tian (Stanford U.)

K. Yoshimura (MSU-Bozeman)

contact info : http://vso1.nascom.nasa.gov/docs/wiki/ContactUs

Virtual Solar Observatory

http://www.virtualsolar.org/

Abstract:

Mission data catalogs are typically built with the specific mission in mind. This can create challenges when trying to abstract the metadata to make it useful to other researchers. The deluge of data from new missions such as STEREO and Hinode have brought in not only issues in scale, but also complexities due to the difference in these new experiments in the context of existing norms.

We will discuss issues and use cases to be considered in designing a mission's data systems in order to better serve the Heliospheric community.

Documentation for Catalogs

Users of the catalogs need to understand what the fields and values within the catalogs mean. As catalogs can be used for cross-discipline work, terminology used in a given catalog may differ from what the scientist using the catalog expects it to mean.

Terms which have no formal scientific definitions should always be documented, so that a scientist can understand how the term is used by the person or team constructing the catalog. Although some values may require significant detail to fully explain, a quick description can allow a scientist to determine if a record is something that they are not interested in, and can later follow up on the items that might be useful.

Examples of Problem Terms:

One catalog uses the term 'Orange' to describe three different filters. Documentation for the instruments reveal that observations using one of the filters are sensitive to a different spectral range than observations using the other two.

Another mission makes the field 'obsdir' available, but does not define what it is, or what all of the valid permutations are. Descriptions of these values were not possible until public interfaces were available and text in the interface was associated with the values in the field.

Accessing the Catalog

Catalogs may be stored such that they can only be accessed through a single software interface. The software may make expectations on how users will expect to interact with the data, and not be useful for general use.

In some cases, it may be useful to have multiple user interfaces to serve each distinct group wishing to use the data. The overhead of maintaining these interfaces can be offset by joining a Virtual Observatory, but the catalogs may need to implement cross-walks to SPASE or other VO interlingua.

In more extreme situations, the data may need to be stored or indexed differently to support queries from another discipline.

Example of a Catalog Structure Problem:

Any catalog that only tracks the concept of 'observation' will be inable to tell anything about the relationship of those observations. Trying to search on observations for a given observing mode where the time cadence is less than 5 minutes may not be possible without fist reprocessing the catalog.

How to Use the Data

As catalogs are designed for a given mission, they typically assume that catalog users are familiar with their data. With the advent of Virtual Observatories and the increase of cross-disciplinary research, it is less likely that someone who finds your data knows how to use it.

Documentation for file formats, calibration and other ancillary files, and software to process the data are as important as the files themselves. Data is not useful if it cannot be understood.

Normalized Metadata

Normalization allows database designers to store their data more compactly by associating attributes through relationships with other entities. For instance, a database might have information about an individual instrument, and then simply store that an observation was taken by that instrument.

Although this can result in savings for storage, it makes it more complex to retrieve information from the catalog if you need information spread across multiple tables. It also cannot account for instrument degredation over time if we assume that the attributes such as spectral range sensitivity are constant for a given sensor.

In some catalog designs, the database may not have information about commonly requested metadata, but it may be inserted or translated by the software used to interact with the catalog. This information must be documented for the catalog to be useful to other parties.

Example Candidate for Denormalization:

All of the following are H-alpha observations in a single catalog:

Camera	Filter Wheel Position	Polarization Wheel Position
0	2	1
1	1	0
2	0	1

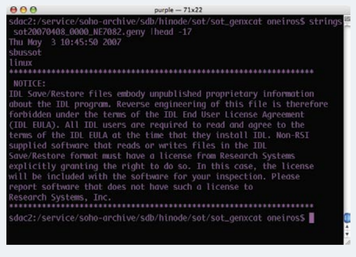
Storing the Catalog

Catalog designers must consider how people may wish to interact with their catalogs. Catalogs may be stored in any number of formats which have different advantages in terms of size, speed of querying and portability.

As data volumes increase, the size of data catalogs increases along with them. Catalogs may be segmented to work around problems with their size, but this can make it more difficult to mine the catalogs for information. As each discipline has different primary attributes for selecting their data, segmented catalogs may prove more problematic for cross discipline research.

The use of some proprietary formats may limit the ability for all scientists to use the data, as they may not be able to import the catalog into the tools they prefer using. Using some formats may require scientists to give up some of their rights to their data, and may have implications for long-term data preservation.

Example of Rights Restrictions:



Provenance

Knowledge of where your data comes from can be important to differentiate between two similar records. Were files from one archive mirrored to another, or were they reprocessed from the same input? This can be a significant issue if one archive was found to have problems with their processing software or if they have made modifications to it.

Tracking provenance may become more of an issue as data is used by other disciplines with more onerous requirements. Organizations that have data that may affect national policy on climate change are beginning to track provenance and handling of all of the inputs into each object in their data pipelines.

What is a Record?

When users search for data, the search system must return some object to the user – but the level of granularity of the object depends on the design of the data catalog.

A mission may catalog multiple formats for a given data object, and may track multiple versions or editions of those objects. As data volumes grow, it becomes more important to make data highly available, but we cannot be assured that each archive serving the data is serving the same data objects based solely on the information in data catalogs.

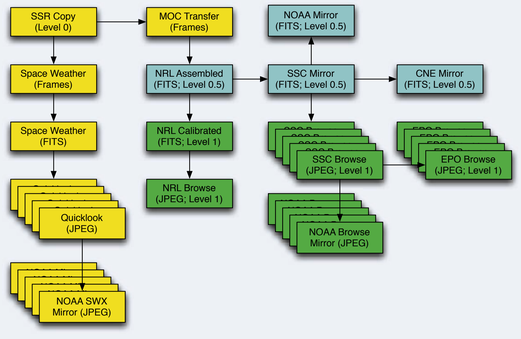
If the raw sensor data was mirrored and then processed independently by each archive, they may not be equally useful for all users. The data may have been processed differently, or they may have different metadata which affects the ability of a scientist to use the data objects with their favorite tools. Packaging may also differ in time granularity, as archives may serve individual observations or hourly collections.

Some data providers may not pre-package their data, and may generate dissemination packages as requested by the users of their systems. This flexibility is very useful to the users, but can make it difficult to merge records from federated search systems.

Each archive can make assumptions about what is best, but the best level of granularity is a function of what it is being used for, not a function of the data itself.

Example:

Without considering the issues of time series data that may be packaged as daily, monthly or for some other time range, we still have issues in describing a single observation, such as an image from a telescope. A single SECCHI observation can come down through two antenna (space weather and via the deep space network), and each is processed into FITS which are processed into multiple quicklook products which are mirrored at multiple sites. Even without considering recalibration, a single SECCHI observation results in 35 data objects, but only 17 distinct ones. Including calibration could result in hundreds of objects tracked for each sensor recording.



Data Models

Data models can be viewed as a constraint on science – if a mission's data can't be fit within a given model, is it a problem with the mission? It is actually a problem with data models, as they cannot predict new and innovative research. Data models may need to be extended to handle data from additional missions.

Consistent data models within a discipline help to make data findable across archives, and benefit re-use of scientific data. Although general research may not be the intent of the given mission, specialty data may still be useful for other purposes.

If it is too difficult to design a catalog to support multiple data models, then don't. Multiple catalogs that are optimized for separate purposes may be preferable.

Identifiers and Foreign Keys

Selection of identifiers is critical to comparing data objects to determine duplication. Unfortunately, identifiers are frequently either based on timestamps, filenames, or dynamically assigned numeric sequences.

Sequences often have no relationship to the data itself, and may not remain consistent if entries are removed from a catalog and later re-inserted. Timestamps are a function of calibration, and can change between level 0 and level 1 data or between different versions of level 1 data. Filenames are frequently a function of processing, and different data products are not directly comperable in catalogs.

Catalogs must be shared when mirroring data to other organizations, so that we can maintain the same keys and identify that the data truly are identical copies.

For the Future

The science community should not assume that their data is cataloged in a way that makes it discoverable to others. Although the Virtual Observatory community can help to define what attributes are useful to make data findable and usable, we cannot expect the mission archives to do the work.

As funding for science declines, we cannot expect each of the individual missions to maintain data modeling experts and documentation experts on their staff. The science informatics community could fill this niche by either defining recomendations and standards, as well as gathering best practices on not only data catalogs but also event and feature catalogs. We must be careful to explain and document why each aspect of the standard is required, and so the mission scientists don't just view them as yet another batch of red tape, but can understand the reasoning behind the requirements.

It may be worthwhile to offer out the informatics community's help on new missions and to get involved in the design of new catalogs, rather than to take a hands-off approach and just assume that we should just wait for the missions to build their catalogs, and integrate them after the fact.