Virtual Observatories and other unifying data systems have been forming in nearly every science discipline. As is common in any field, language evolves to discuss the concepts, but it may evolve differently when communities don’t intercommunicate. In order to discuss our organizations and data systems across disciplines, we must have a clear language to be able to communicate information about our systems and the concepts within our systems. We present common terms and definitions used in earth and space informatics when discussing science archives, search systems, services and other data system components. One benefit of a common vocabulary is to help those who implement science data systems to easily recognize other efforts with a common purpose. A common vocabulary is also useful in identifying analogous terms in other fields such as computer science and science and information science. This is an update based on feedback at the AGU 2010 Fall Meeting, RDAP 2011 & 2012, HDMC 2011, online discussions and scientist interviews. [Changes since the previous release are noted in red text. Contact Joe for alternative formats for Section 508]

Other versions at: http://virtualsolar.org/vocab

‘Terms’ vs. ‘Concepts’:

When discussing vocabularies, it is important to remember that a term is simply a word or phrase that refers to a concept. The concept has a definition. We often have cases where two fields share the same concepts, but may use different terms to refer to them; in other cases, we might have the same term used for similar concepts (eg. ‘sample’), or even completely unrelated concepts (eg. ‘granule’). These terms are defined as they are commonly used in science or scientific data systems. We are mostly focused on concepts unique to the scientific and space data domains and terms that are necessary for information exchange and operation. Some terms may be used inconsistently as compared to other fields or even across scientific disciplines.

As you read this, keep the following in mind:

We are looking for problematic terms that people have encountered, particularly where there is lack of agreement on the definition between various disciplines. Are we missing any terms necessary to discuss science data systems? Do you disagree with any of these definitions? Have you encountered any other problem terms? If so, tell us. (see e-mail addresses to the left) … and include what field/discipline you’re in.

What is ‘Data’?

There are a few different philosophies on the nature of data. Both computer science and information science use the term data to describe a much broader concept than most scientists would qualify as data. Even within the physical sciences, there is disagreement about if data comes only from observation or if the results from computer models are included. To be as inclusive as possible, we will consider data to be any record which can be used to support a scientific argument, even if it may not be considered valid evidence in all disciplines. In the social sciences, data may include survey responses, interviews and historical documents.

Data Systems

There are many terms used to describe the systems used to store scientific data. Data systems are any system built to store or manage data; it may be a simple file structure accessible via FTP or a more complex system involving databases and search. The term data system typically implies digital data, and is not used for physical recordings.

Catalog
A list of some sort; it may be a stored in flat file, a series of files, or within a database.

Data Catalog
A list of data objects available within a data system. It may only contain paths to the files stored, or it may contain additional metadata for searching.

Union Catalog
A merged catalog from multiple data systems.

Federated Search System
A search interface that searches across multiple data systems, and provides a merged result. The queries may be sent out to the other data systems in real time (distributed federated search), make use of a union catalog or local mirrors of data.

Virtual Observatory (VO)
A federated search system for scientific data; typically they focus on data from one scientific discipline (VxO).

Registry
A type of catalog in which metadata is recorded about some set of objects. This term is particularly confusing as the final VO archive hands the data off to the permanent archive once it is no longer being used by the community. It may also serve as a mirror for a resident archive.

Data Archive
A repository with limited or no user access to data objects.

Permanant Archive; Long-Term Archive
Archives in which the primary goal is preservation of the data rather than immediate usage.

Deep Archive
A permanent, dark archive.

Rolling Archive
Archives that only keep the most recent data; older data is purged to make space available for more recent data.

Types of Archives

Whereas repositories tend to be storage systems, the term archive is commonly used for the organization that manages the storage of objects and attempts to preserve or curate the data objects. Note that data may exist in more than one archive; a deep archive may be the backup for a resident archive.

Data Archive
An organization that manages one or more data repositories. Sometimes called a science archive, but that term may be ambiguous as it may be an archive of documentation, reports and journals.

Active Archive
A data archive in which the data being collected are still being actively updated; it may grow with time, or still be undergoing calibration. As such, identifiers to data may not point to a specific edition, so they do not qualify as an archive to the science community. It may collect data from a single team or investigation (Instrument archive), or from all of the investigations from a larger project (mission archive, project archive).

Resident Archive
An archive in which the primary goal is easy accessibility and usage of the data; it will typically have scientists available who can assist in using the data.

Final Archive
The archive that will be responsible for the data after an investigation ends, but while the data are still being used by the science community. This term is particularly confusing as the final VO archive hands the data off to the permanent archive once it is no longer being used by the community. It may also serve as a mirror for a resident archive.

Virtual Observatories and Data Systems (v2.3)

(2012- November-30)

J. A. Hourčé
(NASA/GSFC; Wyle; VSO)
T. A. King
(UCLA/GPP; VMO; PDS)
joseph.a.hourcle@nasa.gov
Tking@igpp.ucla.edu
Data Granule
A grouping of data objects that is the smallest amount individually addressable and retrievable within a given data system. As a value in the scientific term in some fields, this concept should always be qualified as data granule; this term is ambiguous when describing amounts of data as it is a function of the data system.

Data Collection
A grouping of one or more data granules.

Metadata
Information about the data being tracked within a data system. Metadata typically conforms to a metadata information model. Metadata may include the name of the sensor or person who collected the data, where the data was collected, information about the units and dimensionality of the data, and other notes recorded by the investigator about how the data has been processed.

Science Metadata
Metadata used by scientists to understand what is recorded in a given file or object, given in standard form for their discipline, which may use accepted terms or physical units.

Engineering Metadata
Metadata used by the investigation team to record information about the sensor and the observation. It may be extensible to scientific use or science metadata may be derived from the values such as filter=d which could resolve to a spectral range or polarization.

Provenance Metadata
Metadata used to describe where something came from; in some fields, this is used to describe the science metadata (location of the observation, name of the sensor) or engineering metadata (details of the sensor's observing model), but it may also include information about how and where the data were calibrated or stored.

Data Processes
Terms used to describe the processes that archives perform on the data it manages. This list does not attempt to enumerate all of the types of processing that may be done by different scientists in their analysis or to reduce the data for dissemination.

Data Curation
Processes done to manage, maintain and validate data to ensure continued access.

Data Transformation
Conversion of values; this may include conversion between different physical units or coordinate systems, or deriving new values from the data.

Data Refactoring
A form of data transformation that typically implies that some analysis, reduction or other irreversible process was applied.

Data Cataloging
Assigning metadata to managed objects to assist in finding or retrieving the data. You should qualify this as science cataloging, in which provenance metadata, identifiers and other administrative metadata are assigned, as some scientists will assume science cataloging and take objection to being performed by people who are non-experts in the data.

Quality Analysis
Review of data objects for consistency, completeness or correctness.

Data Harmonization
Transformation of heterogeneous data to a common reference frame or otherwise normalized.

Data Policies
Rules applied to the data; this may be legal (use, distribution) or administrative (number of copies maintained, procedures for ingestion, frequency of verification).

Use
Has a number of different meanings; some groups differentiate between using the data for its originally collated purpose and data re-use or data repurposing as using the data for a different purpose. The vagueness of data use may cause a number of problems when using creative commons licenses for data, as it is unstated what qualifies; if publishing research using the data would violate ‘no derivatives’ or if publishing in a non-open journal would violate ‘no commercial’.

Data Distribution
Sharing the data in its original form while data re-distribution is sharing of the data in some form other than what was provided by the original archive. Different groups draw the boundary of ‘some other form’ differently; refactoring (eg, data that has been included, but repackaging (even when aggregating with data from other sources) or accepted reversible transformations may not be.

Qualifiers for Data
Raw Data; Unprocessed Data Values emitted from a sensor.
Calibrated Data Data with known sensor effects removed
Corrected Data Data with noise and other problems removed
Reduced Data Data that have been transformed to a lower resolution than the original data; this includes daily or weekly averages from time series or 2x2 binned images.
Compressed Data Data that stored in less space on disk. Includes non-lossy compression, where the original values are recoverable, and lossy compression, where precision is lost.
Processed Data Data that store some value or property in the data. Includes removing trending, where the original values are recoverable, and removing trending, where precision is lost.

Results Data
Donates that it is not raw data, but says little about the type of processing that has been applied.

High Level Data
Data that it is not raw data, but says little about the type of processing that has been applied.

Derived Data
Data that is not raw data, but says little about the type of processing that has been applied.

High Level Data
Data that is not raw data, but says little about the type of processing that has been applied.

Derived Data
Data that is not raw data, but says little about the type of processing that has been applied.

High Level Data
Data that is not raw data, but says little about the type of processing that has been applied.

Derived Data
Data that is not raw data, but says little about the type of processing that has been applied.

High Level Data
Data that is not raw data, but says little about the type of processing that has been applied.

Derived Data
Some systems will describe the relationships between data objects, and these terms mean different types of processing to different objects.

Some systems will describe the relationships between data objects, and these terms mean different types of processing to different objects.

Some systems will describe the relationships between data objects, and these terms mean different types of processing to different objects.

Some systems will describe the relationships between data objects, and these terms mean different types of processing to different objects.

Some systems will describe the relationships between data objects, and these terms mean different types of processing to different objects.

Some systems will describe the relationships between data objects, and these terms mean different types of processing to different objects.