

Vocabulary for Virtual Observatories and Data Systems (v2)

J. A. Hourclé
(NASA/GSFC; Wyle; VSO)
joseph.a.hourcle@nasa.gov

T. A. King
(UCLA/IGPP; VMO; PDS)
tking@igpp.ucla.edu

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 content 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 information science.

We solicit input on problematic terms that people have encountered, particularly where there is lack of agreement on the definition between various disciplines.

This is an update based on feedback at the AGU 2010 Fall Meeting, online discussions and scientist interviews.

'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 science data systems and terms that are necessary for interoperating and that may be used inconsistently as compared to other fields or even across scientific disciplines.

As you read this, keep the following questions in mind:

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 above)

... 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.

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 flat file, a series of files, or stored 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 catalogs** (**local federated search**).

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 sort of object. Their usage varies greatly to include **data registries**, recording the existence of **repositories** or other **data collections**; **event registries**, recording the properties of events seen within the **data**, or **vocabulary registries**, recording **terms** and their definitions as used by a **controlled vocabulary**. Registries differ from catalogs in that they have a mechanism for accepting entries from more than one source.

Repository

A system for storing objects; the objects typically cannot be interacted with from within the repository, and must be retrieved to be used; it may contain digital or physical resources. This includes **institutional repositories**, which collect articles, documents and other research or output from a university, company or other institution; **discipline repositories** that collect

resources for a given discipline or group, such as arXiv or ADS; or **data repositories** that focus on collecting **scientific data**.

Types of Archives

Whereas **repositories** tend to be storage systems, the term **archive** is commonly used for the organization that manages that storage and attempts to **preserve** or **curate** the **data objects**.

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 or journal articles.

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 library community. They 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 community. This term is particularly confusing as the **final archive** hands the data off to the **permanent archive** once it is no longer being used by the community.

Permanent Archive ; Deep Archive ; Long-Term Archive

Archives in which the primary goal is **preservation** of the data rather than immediate usage.

Rolling Archive

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

Objects and Entities

Data systems store a number of different things, grouped in whatever manner makes sense for that system's intended purpose. The following terms are necessary for describing what is stored or described within a **data system**.

Resource

Anything that can be given an identifier and referenced; this includes **data**, documents, **data systems**, people, sensors and projects.

Science Data

Values collected as part of a scientific investigation; these values have a reference frame which may include coordinates, units,

accuracy or precision. It is important to use the full term, as *data* has incompatible definitions in other fields, such as computer science (anything encoded as bits) and information science (any encoded information).

Observational Data

Science data collected through observation. This includes uncalibrated values (**raw data**), derived values (**calibrated data**), and other transformations of the values (**processed data**). This includes *experimental data*, in which the scientist creates the events being observed.

Simulation Output

Values collected as the result of computer simulation or other modeling. As there are a number of fields that do not consider these values to be *science data*, it is best to separately qualify it. The term *simulated data* should be avoided, as it may refer to fake *data objects* created for testing; the term *model output* should be clarified if used, as *model* may refer to a number of different concepts; and the term *computational data* is ambiguous as it may refer to input or output from a simulation.

Data Object

A grouping of *data* and its associated *metadata*. Typically the *data* is from one observation, but the concept of observation varies by discipline; it may be a single scalar sensor recording, vector quantity, multiple coordinated readings taken at the same time, an image or something else entirely.

Data Granule

A grouping of *data objects* that is the smallest amount individually addressable and retrievable within a given *data system*. As *granule* is a 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 extraneous to scientific uses or it may be derivable to *science metadata*, such as 'filter=4' 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 mode), but it can also include information about how and where the data were calibrated or stored.

Please note that the terms *dataset*, *data product* and *data series* are not defined here. Although all refer to a grouping of *data granules*, the terms are used inconsistently across disciplines; in solar physics, a *dataset* is a collection of *data products* while in earth sciences, a *data product* is either a collection of similar *datasets* or a classification of *datasets*. These terms should be avoided, or clearly defined when used.

Some systems will describe the relationships between *data objects* as being of a different *version* or *edition*, however these terms mean different types of processing to different communities, and should be clearly defined.

Vocabulary Systems

To ensure consistent use of metadata terms in scientific data systems, there are a number ways of managing those terms; although these terms may not be widely used in the science fields, their inclusion is necessary to be able to discuss our data systems to achieve interoperability.

Controlled Vocabulary

A prescribed list of *terms*, each one having an assigned, unambiguous, non-redundant meaning.

Taxonomy

A *controlled vocabulary* organized into a hierarchy.

Thesaurus

A *controlled vocabulary* in which *concepts* are represented by preferred terms, organized into one or more hierarchies and that includes synonyms and other equivalent relationships, and may include notes and instructions on usage of terms or relationships between related concepts. Note that *Roget's Thesaurus* is a *synonym ring*, and not a *thesaurus*.

Ontology

A *controlled vocabulary* in which any number of relationships can be defined; it may include equivalence and hierarchy as in a thesaurus, but can also include relationships for reasoning. (eg, 'fishA' lives in 'regionB'; 'fishA' eats 'fishB'; 'telescopes' generate 'image data')

Crosswalks

Linkages between *terms* in two different *controlled vocabularies*.

Folksonomy ; Free Tagging

Used to describe vocabularies without any formal control; terms may be ambiguous or redundant.

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 to assist scientists in their analysis or to reduce the data for *dissemination*.

Preservation

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

Data Curation

Processes done to maintain and add value to data for both current and future use. This includes *preservation*, annotating data or improving the documentation.

Ingestion

Processes done on receipt of data to insert it into a *data repository*; this may include validation or other quality analysis, *transformation* of the data or metadata, *cataloging* or *repackaging*.

Dissemination

Processes done to distribute the data from the repository to some other person or system; this may include *transformation* or *repackaging*.

Transformation

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

Cataloging

Assigning *metadata* to managed objects to assist in finding or retrieving the data. You should qualify this as *science cataloging*, in which events, features and other phenomena are identified and classified vs. *data 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.

Media Refresh

Transferring the data from one physical storage media to another, to prevent data loss from failure of physical media or lack of readability due to changing technologies.

Repackaging

Moving the data between different digital containers; this may include converting between different file formats, aggregating a series of values into one file, or creating a tar, zip or similar archive file. In some cases this may also include *transformation* of the metadata included with the data, but there is the assumption that the values for the data remain the same.