Coordination Action for the integration of Solar System Infrastructures and Science

Project No.: 261618
Call: FP7-INFRA-2010-2

Report on how to enhance the Metadata

Version 1.0

Title:	**Report on how to enhance the Metadata**
Document No.:	CASSIS *Deliverable:* **D2.2b**
Date:	20 November 2013
Editor:	**R.D. Bentley**, UCL-MSSL
Contributors:	J. Aboudarham (OBSPARIS)
Distribution:	Unrestricted

Revision History

Version	Date	Released by	Detail
0.1	15-Mar-2012	R.D. Bentley	Initial version
0.2	20-Apr-2012	R.D. Bentley	Rework after working on Use Cases
0.3	10-May-2012	R.D. Bentley	Tidying, reorganize
0.5	30-May-2012	R.D. Bentley	Working version; submitted
0.9	20-Oct-2013	R.D. Bentley	Heavily revised
1.0	20-Nov-2013	R.D. Bentley	Released version

Need more information?

For information about this deliverable please contact Bob Bentley (b.bentley@ucl.ac.uk)

Contents

Overview

The purpose of this document is to examine how improvements can be made to the different types of metadata used within projects in order to increase interoperability between related scientific domains.

The document is very much work in progress. Initially the study mostly contained ideas gained from the experience in working on the HELIO, SOTERIA and Europlanet RI projects but there is a desire to generalize the scope in order to consider Solar System Science as whole.

Our interaction with user communities has led to an agreement to provide guidance material for ESA'a Solar Orbiter project, scheduled for launch in 2017/18. This project needs to develop a data system that will be used to exploit data from the mission and we are viewing this as a way to examine the validity of our ideas and developing them further. From a Solar Orbiter perspective, there is a need to have proposals for key metadata available during 2013.

Science Use Cases gather in another CASSIS Deliverable (D3.1) have helped widen the scope of what we are looking and work on metadata from other domains will start in June 2012.

Together with Deliverable D2.1 (Survey of Existing Services), this deliverable provides information that will be used to prepare the recommendations for the community (Deliverable D2.3).

Note that CASSIS is a Coordination Action and there are therefore limits of how far we can develop the ideas within the scope of the project. Some prioritization is therefore necessary – we cannot do everything – but working with projects such as Solar Orbiter will increase exposure of CASSIS and held us realise our objectives.

Reference Material

This deliverable is drawing on expertise built up in other projects and other bodies, including:

- Data Models from HELIO*, Europlanet RI IDIS* and EGSO
- HELIO Deliverable N3.3* (Metadata Standards)
- Space Physics Archive Search and Extract (SPASE) Data Model
- Standards developed by the International Virtual Observatory Alliance (IVOA)
- Various standards developed by the International Standards Organisation (ISO)
- Dublin Core (with some extensions)

Note that some of the sources of information have only recently become available (*).

Introduction

There is an increased desire to do science that spans existing disciplinary boundaries. Since advances in technology mean that providing access to archives of data and to processing and other capabilities is now largely a matter of plumbing, the data and metadata are becoming the greatest impediment to facilitating science. The communities have evolved independently and differences in the way that they store and use their data are significant issues; the problem is further exacerbated by the rapid growth in volumes of data making it increasingly difficult to identify things of interest by simply trawling through the observations.

Metadata is a key tool in an environment where the desire is to provide integrated access to a wide variety of data sources and services across several domains. It is important that the metadata are as complete, accurate and self-contained as possible and are compliant with an agreed set of standards.

The ability to identify time intervals and locations of interest based on different types of metadata and derived data products is becoming increasingly important as the volumes of data increase. The search may be based on the occurrence of events, features and other criteria; this is support by other types of metadata that facilitate access to services necessary to exploit the data. The search capability is supported by other types of metadata that allow the system to function by facilitating access to data providers and other types of services.

Improving the quality of all kinds of metadata it is therefore a cornerstone in the drive to improve interoperability and in this deliverable we examine how this can be done through the adoption of standards.

Although initially based on the three projects involved in CASSIS we are trying to look beyond these and make the discussion as agnostic as possible. The science Use Cases gathered in Deliverable D3.1 have helped widen the scope; they have convince us that we need to think of the interoperability that we would like to achieve within a very broad collaborative research environment that could be formed from a series of overlapping environments targeted at specific areas.

We are considering the Solar System Science as whole – i.e. a wider remit than is covered by the individual projects. While the projects involved in CASSIS might adopt standards developed by bodies such as the IVOA (and to a lesser extent the IPDA). If we include related communities such as geoscience then we also need to consider standards from organizations that are relevant to them. In all cases, any changes in standards that we find are needed to facilitate interoperability will be communicated back to these organizations with the aim of gradually broadening the scope of such standards.

The objective of Deliverable D2.2 is to identify ways of enhancing the metadata in number of areas. Together with Deliverable D2.1 (Survey of Existing Services), this deliverable provides information that will be used to prepare the recommendations for the community in Deliverable D2.3

In the next sections we review the types of relevant metadata and then study each in detail. We also examine whether drawing up list of common terms in Observational Metadata based on data from the three projects would facilitate interoperability and look at standard terms that can be used for event data.

Types of Metadata

The metadata can be divided into a number of areas although the delineation is not necessarily unique and might be grouped differently elsewhere. For the purposes of this document we have chosen:

- **Descriptive Metadata** – Describes a resource for purposes such as discovery and identification. There are many types of descriptive metadata but here we will highlight the following:
    - **Observational Metadata** – Related to the way in which a given piece of Data was obtained or processed.
    - **Derived Metadata** – Extracted from Primary Data through subsequent analysis or processing.
    - **Search Metadata** – Defines the criteria and boundaries of the search. Also includes information that is useful to facilitate the search.
- **Structural Metadata** – The organizational characteristics of the Data including how it relates to other Data
- **Administrative Metadata** – The characteristics related to methods of access and management of Data or other Resources

A **Data Model** is used to describe the relationships among the metadata elements.

Which metadata are needed depends on what you are trying to do and the complexity of the system:

- If the desire is just to produce data then a users' interest is mainly in the *Observational* and *Derived Metadata* parts of this document.
- When it is decided to provide a capability to distribute data or provide some other functionality then the interest also encompasses the *Structure Metadata*.
- If the capability is part of a larger system then the *Administrative Metadata* becomes important.

The *Derived Metadata* is potentially the easiest to change since it generated by processing the data; of course it can also be the most diverse.

Descriptive Metadata

General Issues

Need for good Metadata

In order to relate data from different domains it is important that metadata should be as complete and accurate as possible and that that it stands by itself with minimal need to refer to auxiliary information.

In order to identify what information is needed it is essential that we think of the bigger picture in considering what the data might be used for. To do this we need to ask:

- Can someone from outside of the experiment team, but within the project, understand the data without assistance
- Can someone from outside the project, but within the domain, understand the data without assistance
- Can someone from a different domain understand and use the data without assistance

Assumed knowledge

One of the problems with a lot of metadata of all types is the amount of assumed knowledge.

An example of this is where something is not properly described because everyone using it knows what it means. For someone that is not familiar with an instrument this can make the data almost impenetrable – they need information that they do not necessarily have access to. The problems can really start when the experiment team disbands at the end of a project and there is no longer anyone to ask.

In order to address this problem care must be taken in defining metadata, particularly observational metadata. Some of this can be addressed through proper annotation of the parameters in order to clearly describe their meaning.

Parameter Annotation

An issue with some file formats – e,g, FITS – is that it is difficult to structure the metadata in a way that completely describes the meaning of the parameters. This problem could easily addressed if the metadata were in XML where it is easy to form constructs that show the relationship of pieces of information. For example, in the VOTable[1] format each parameters is defined in a FIELD record that contains many different clauses.

The annotation of parameters, and a way in which existing file metadata could be enhanced using an associated external XML file, are discussed in Appendix B.

[1] VOTable is a format defined by the IVOA – http://www.ivoa.net/documents/VOTable/

Coordinate Choice

How to describe coordinates is important in *Observational Metadata* but could also be important in other parts of a data system, particularly within *Derived Metadata*. The choice of spatial coordinates depends on what is being measured; temporal and spectral coordinates should not have the same issues.

Spatial coordinates

Given the diversity of the different types of science within the Solar System it is not possible to define a single coordinate system than can be used in all cases. The coordinate system that is most suitable depends on the domain, the science topic and Solar System body in question.

For example, when undertaking a study in geo-sciences the coordinate system needs to be related to the Earth. If the study is of things on or near the surface, a coordinate system related to the rotation axis of the Earth may be appropriate; if the study is of the ionosphere or magnetosphere then the coordinate system that is related to the Earth's magnetic field is required. The further from the Earth's surface, the more the coordinate system needs to be related to the Earth-Sun line.

In some cases observatories might need to be described in two different coordinates systems. The location of missions at L1, such as ACE, might be described in a coordinate system based on the Sun for a study about the effects of the Sun on the heliosphere, but in a coordinate system based on the Earth for a study of the effects of the Sun on the Earth's environment.

In order to maximize interoperability we need to establish an agreed but limited number of related coordinates systems[2] that can be used as required. The HELIO project produced a document describing the *Spatial Coordinate Systems*[3] that it is using for different part of the problem. This can be used as the basis for the discussion.

The IVOA has produced a set of recommendation on Space-Time Coordinate Metadata[4] for the Virtual Observatory that contains some useful material. It is more oriented towards astrophysics but a few additions and modifications should correct this.

Although oriented towards astrophysics, the WCS has a relevant standard: *Representations of world coordinates in FITS* – http://adsabs.harvard.edu/abs/2002A%26A...395.1061G. More relevant to solar physics is *Coordinate systems for solar image data* by W.T. Thompson (NASA/GSFC) – http://adsabs.harvard.edu/abs/2006A%26A...449..791T

Neither of the standards deals specifically with in-situ measurements where the choice of coordinates is much more complex. This type of measurement is needed within the heliosphere and planetary environments and is not a normal part of astrophysics.

---

[2] At times in the past, new coordinate systems were developed as new areas of science were opening up; this has led to systems that may very similar or even identical but use a different set of names or units. Some rationalization of coordinates is therefore required.

[3] HELIO Coordinates – http://www.helio-vo.eu/documents/public/HELIO_Coordinates_100322.pdf

[4] IVOA STC document – http://www.ivoa.net/Documents/latest/STC.html

Temporal coordinates

Date and Time should be **Coordinated Universal Time** (**UTC**) or one of its components such as International Atomic Time (TAI).

In some datasets, time is related to a particular epoch – in the case of a spacecraft this might be launch, start of the cruise phase, a planetary flyby, orbital insertion or some things similar. This is not a problem so long as the primary time is in UTC and the relative time is provided as an *additional parameter*; if it replaces the proper time information then the data can be difficult to interpret for anyone that is not familiar with the data (c.f. assumed knowledge).

The same argument holds where dates are given as day of year: so long as this is supplied as an *additional parameter* there is no problem; if the information is mixed with the other elements in the time field then the temporal information becomes very difficult to assimilate in any automated fashion.

While the selection of appropriate spatial coordinates depends very much on what is being observed and the location of the observer, there ought to be fewer issues in relation to how time is expressed. However, for data sets that are not generated by commuters, the formats used for the time fields are proving to be very variable, particularly for event data.

In order to make it easier to correlate observations from different instruments and domains, time should be expressed in a **format** that does not require translation. An appropriate format has been developed by the **Consultative Committee for Space Data Systems** (**CCSDS**[5]) and experiment teams should be encouraged to use this in all places that time needs to be expressed.

There is a draft[6] WCS proposal: *Representations of Time Coordinates in FITS.* An ISO standard also exists: *ISO 8601 – An International Standard for Date and Time Formats*

Spectral Coordinates

Depending on the type of observation and instrument, spectral information might be expressed in wavelength, frequency or energy.

The units being used should always be properly described and as far as possible standard factors of 10 should be used; ensuring that this is done is one of the difficulties associated defining metadata (c.f. parameter annotation in Appendix B).

Confusion can occur when the orders of magnitude associated with a value are different for different communities. For example, traditionally there is a decametric radio community in solar physics but a decimetre is no longer a normal fraction of a metre and another community may use a different term.

There is a relevant WCS standard: *Representations of spectral coordinates in FITS*
http://adsabs.harvard.edu/abs/2006A%26A...446..747G

[5] CCSDS – http://public.ccsds.org/default.aspx

[6] WCS draft at http://hea-www.cfa.harvard.edu/~arots/TimeWCS/WCSPaperV0.982.pdf

Observational Metadata

The Observational Metadata describes the way in which a given piece of data was obtained or processed.

It is important that the observational metadata should be as complete and accurate as possible; in addition, for maximum interoperability the observational metadata should stand by itself with minimal need to refer to auxiliary information. In some circumstances, if the metadata are not properly formed it may not be possible to use the observations.

Even if the data are from very different domains, there are certain pieces of information that always need to be conveyed – date and time of the observation, name of the observatory and instrument, etc. There are also parts of the file that contain information that is specific to an instrument or domain.

The *key to improved interoperability is to maximize the number of common parameters used in the observational metadata* – that is, push the boundary between the common part and domain-specific part as far down as possible.

The observational metadata is largely defined when a data system is established for a new instrument or observatory. It is extremely important that a lot of thought goes into the design of this since making changes can be difficult once the data train has started rolling.

Note that two locations are required for observational data: one describes the location of the observatory (this may refer to an orbit ephemeris file for spacecraft); the other describes the volume in coordinate space that is occupied by the observational data.

It should also be noted that a move to use standards to describe the observational data has implication on the *Derived Metadata* since a significant part of this is based in the *Observational Metadata*.

Common Part

The common part certain pieces of information that always need to be conveyed – date and time of the observation, name of the observatory and instrument, observing domain, location of the observatory and the target of the observations, etc.

In order to maximize interoperability the names, meaning and possible contents of parameters used in the common area need to be agreed by as wide a community as possible – i.e. a standard needs to be established. See Appendix B for suggestions in this area.

Instrument/Domain Specific Part

There will always be large parts of the metadata in a file containing observations that is specific to the instrument or domain.

Any attempt to standardize this over too wide a community would be counter-productive but *within a domain it is important that discussions take place and a set of parameter names and their meaning agreed* – these should be described within some form of data model.

Derived Metadata

Much of the information useful for queries is locked away in the data themselves:

- What type of observations was being made
- Where were the sunspots?
- When were there flares?
- How strong was the spectral line?

The information must be extracted from the observations; this derived metadata can be stored as additional catalogues that are used in the search and analysis process.

HELIO has created a number of catalogues that fall into this category and we will examine this to consider what needs to be done.

Event Lists

Events lists are generated from time series data or a set of images – an event is basically when a change of some kind is observed. It is difficult to consider any event list as definitive – each list is created by comparing the data to criteria defined by an observer and a variety of criteria can be considered to be equally valid. It is therefore essential that the purpose of a list and the criteria used are known and understood; it is also useful is the common parts of any lists are sufficiently similar that they can be easily compared.

Event lists ought to be relatively straight forward if they relate to a single observatory but this would require collaboration that may occur but is not consistently adopted/implemented.

Gaps in the data can lead to confusion if there is no way knowing that these exist; also instruments can saturate or even stop working during certain types of events – see Observational catalogues.

What we have learnt from HELIO

The HELIO Heliophysics Event Catalogue (HEC) contains nearly 70 event lists from a number of sources and covering several domains. In order to make it easier to use the information in the lists and compare lists from different sources with each other we have found it necessary to *condition* the data as it is ingested.

Semantics and on-the-fly conversion might appropriate for observational data but for event lists it was decided that it would be more efficient to condition the data once as it was loaded. Conditioning of the metadata also allows us to ensure that parameter names are used consistently and avoid the unnecessary use of synonyms.

Where a researcher has compiled lists where the information from several observatories is combined together it becomes more difficult to describe – there can be multiple starts and stops, possibly multiple locations. (Addressing this is still work in progress for HELIO).

Some of the things that have had to be done include:

- Time formats – (ISO 8601: An International Standard for Date and Time Formats)
- Pointing, location
    - Standard set off coordinate systems, incl. near planets
    - Should the observatory location be included – if not, assumed knowledge...
- Field of view, plate scale
- Names of terms and contents (data model, UCD and *utypes* from the IVOA)

Event Data used on other projects

In a number of areas HELIO has adopted standards developed by the IVOA although extensions or modifications have been needed in some cases.

HELIO has tried to be sympathetic to the standards developed is for event lists (VOEvent[7]) but as encountered difficulties because of the variety of lists that have been ingested and because the standard is not totally applicable to science inside the Solar System.

The Heliophysics Event Catalogue and Heliophysics Events Knowledgebase (HEK[8]) developed by Lockheed Martin for the Solar Dynamic Observatory (SDO) is based on VOEvents but is should be notes that there are differences between how HELIO and the HEK handle event and feature data:

- The HEK considers everything to be an event, including feature information derived from images.
- HELIO stores event list and feature data derived from processing images in different databases. If a feature appears or disappears then it counts as an event but otherwise features are just records in the feature catalogue – see below.

The two approaches both have merits. It is easy to find collocated "events" in the HEK although the size of the database could become an issue; the HEC and HFC are better able to hand the different type of event and feature data.

The Heliophysics Data and Model Consortium (HDMC[9]) has produced a *Specification for a Heliophysics Event List Format*[10]. The Heliophysics Event List format is a plain text data format with metadata embedded in comments. The HDMC claims that the metadata is sufficient to properly interpret typical event lists and can point to other sources of information related to the data. The format is designed to be easily parsed and to be easily converted to a VOTable format.

For maximum interoperability we would need to consider event data when we extend the discussion to the related domains – for example, what constitutes an event in the geo-sciences.

Feature Lists

The HELIO project has create the Heliophysics Feature Catalogue (HFC) by processing images from various sources to look for different types of solar and heliospheric features; the HFC was based on the Solar Feature Catalogue (SFC) that was developed several years before by the EGSO project.

A detailed description of features was developed for HELIO but it does not as yet represent a community-wide standard.

---

[7] IVOA Document VOEvents – http://www.ivoa.net/documents/VOEvent/

[8] The HEK was developed by Lockheed-Martin for the NASA's Solar Dynamic Observatory project – http://www.lmsal.com/hek/

[9] The HDMC was created to provide coherence and long-term support for a number of aspect NASA's Heliophysics Data Environment, including VxOs and related services, etc.

[10] http://www.spase-group.org/docs/conventions/HDMC-Event-List-Specification-v1.0.pdf

Observational Catalogues

An observational catalogue is a summary of the observations made by an instrument – such a catalogue would normally be derived from the metadata associated with the observations (i.e., in the file headers). Here we differentiate between an *observational catalogue* and an *observing log* – the former is generated from the observational data while the latter should be a record created by the observer as the observations were being made (see the section about *Observational Metadata*).

One problem is that only a few instruments create such catalogues. While in principle it is not difficult to generate them, the task should ideally be undertaken by the experiment team, or a group that knows the data and has access to a complete copy of the data.

If the catalogues are generated from partial copies of the data, this can result in confusion when they are used in the search for observations. Similarly, creating the catalogues "on the fly" from that data that are available can give a biased perspective; the site being accessed may not have a complete or up to date copy of the data and the catalogue would only contain what had been found not what observations were made.

This topic touches on to the issue of **data provenance**, a subject that has not been handled properly (if at all) in the past and must be addressed for future datasets – see under *Administrative Metadata*.

In HELIO, the Unified Observing Catalogue (UOC) is used to address inadequacies in the catalogues that are available for some instruments and also address some complexities related to data access for some datasets. The UOC contains a number of tables, including:

- Pointing (and other mode) information for limited field-of-view remote-sensed solar instruments. These are abstractions of the observational catalogues from the instruments ingested using a standardized set of parameters.
- Planetary data in NASA's Planetary Data System (PDS) and ESA's Planetary Science Archive (PSA). The archives are difficult to access; they have been harvested to determine the location of the files and the results are stored in the UOC.
- Other information that it is hard to discover by conventional means

As notes, in creating the tables loaded in the UOC it was again necessary to standardize the parameter names and units used.

Engineering Logs

Data coverage is rarely 100% and there can be many reasons why observations were not made during a particular time interval – by design, planned or unplanned (engineering) downtime of the instrument or observatory, loss of telemetry, radiation belt passage (spacecraft), weather (terrestrial), etc.

After a project has ended, and even while it is active, it can be very difficult to determine why observations are not available unless information describing this type of occurrence is properly recorded. If this is not done there is a danger that a scientific user may jump to the wrong conclusion about why nothing was seen – just because you do not have observations does not mean that an event did not happen.

It is therefore essential to keep records – in electronic form – that describe all times that the instruments were not operating normally.

Unfortunately at the moment instrument teams do not always create engineering logs but the need to do this should be strongly emphasized. It is desirable that the requirement for this type of logging is built into the data system from the outset since it involves information that it

may be difficult to reconstruct even relatively shortly after anomalies have occurred; the system should flag any time that a unexpected gap in the data has occurred.

At a ground-based observatory there will often be a permanent team responsible for operating the telescope; this team may set the telescope up before handing over to the "visiting" observer. As a matter of course the team would records details of the configuration, weather and viewing, etc. in an observing log; along with this would be information if anything out of the ordinary happened.

Space-base observatories generally operated differently. Sets of command that represent the observing plans can be generated hours or even days before they are executed and there is often little chance to make adjustments in real time; it can then take days to retrieve the data with variable delays depending on the ground stations used. The consequence is that things become decoupled; there are often a different people responsible for planning, executing and monitoring the data and some actions take place long after the observations were planned.

While this may seem a trivial problem, the consequence is that operational issues are not always properly recorded and even a short time after the observations were made if can be difficult to establish why something did or did not happen. This can cause particular problems for people external to a project especially those from a different domain or when someone is looking at the data after the project is finished and instrument personnel are not longer available.

Search Metadata

Search Specification

In conducting a search, various pieces of information are used to identify a phenomenon that the researcher wishes to study. The information might include event and/or feature lists and other derived products, but could be based on some other criteria.

The result of a search is the information needed to source particular types of observations at specific location and times. If the phenomena evolve over time, and affects a series of locations, then some form of model is required to connect these and identify when and where effects will be observed. Using metadata that helps match instruments to locations and times, the search is reduced to a list of time intervals for specific instruments (the location is implicit).

To start a search it is first necessary to define the criteria and bounds of the search – these help describe how the system needs to access the derived metadata and observational metadata.

It is difficult to completely generalize the criteria since they depend on the science use cases that need to be supported and the types of data that need to be addressed. Generally the criteria should include the specification of:

- A time interval of interest and a special region of interest
- The type of phenomena that it being studies, where (and when)
- The types of observations needed to observe the phenomena (with parameters narrowing the choice)

Other Search Metadata

There is another set of *Derived Metadata* that consists of consolidated or aggregated metadata from a number of sources.

In this type of metadata, information are pulled together so that they are easier to use – sometimes this type of metadata relates specifically to the project if formulated correctly but could be useful elsewhere. Examples of this in HELIO are the Instrument Capabilities Service (ICS) and the Unified Observing Catalogue (UOC). The Instrument Location Service (ILS) also serves this function but potentially has wider applications.

Structural Metadata

Structural Metadata describes the actions that can be performed on data and metadata elements. It also allows the use of software repositories and processing services to perform actions on data in a well-defined manner. In essence it is the verbs to other metadata's nouns.

The *Structural Metadata* should be designed with change in mind. During the lifetime of a large project there will probably be a number of changes in technology and the way that systems interact and the design of the system metadata should not preclude being able to benefit from these changes. It therefore needs to be modular and layered, with all required information being contained in files that can easily be extended and modified.

It is also important that the design should be try to be agnostic of the domain as far as possible – this will facilitate interoperability.

File Naming and Storage

Assuming that by improving the quality of the Observational Metadata the data can be made easier to use across domain boundaries, the next task is to make them more accessible. Following a few simple rules in the way files are named and in the directory structure used to hold them can make a significant difference to how easily they can be accessed.

There are no hard and fast rules for file names, but the names should be sufficiently unique that files can be stored outside of their "native environment" – i.e. where they are normally stored on the system of the group that generated the file. In other words, the file should be able to exist without causing confusion when removed from the context of where it is normally stored.

The file names should also ideally identify the type and origin of file, the nature of the data and when the observations were made; the name might also indicate if an image represented a partial or the full field-of-view of the instrument.

Although not strictly metadata in itself, the structure of the archive can make a lot of difference to how easily the data can be accessed. If the data are held in a hierarchical structure based on date and time, it is much simpler to create metadata that can be used to find the required data. Storage with this type of structure is essential for resource-poor providers and would be beneficial for data centres.

A summary of observations that have been made, or *Observational Catalogue*, is useful and simplifies access. It is particularly useful if not all the observations in the archive are available on-line.

Administrative Metadata

Administrative Metadata is used to locate and access resources within the system. Two of the main elements are the resource registry and a means of authenticating and authorizing users

In principle the *Administrative Metadata* could be and should be domain independent; this will simplify interoperability with Virtual Observatory projects from related communities. However, the management of Administrative Metadata across multiple collaborative environments needs to be discussed.

Data Management

File Formats

It is not appropriate to force all varieties of data into a single file type – some types of file are more appropriate for certain types of data than others and research infrastructures and VOs should be able to accommodate all types of data are available.

However, there is a problem with many of the existing file formats that they do not properly define what parameters in the file represent. For example, the FITS standard is used widely in many astronomical fields but is actually quite a loose standard. There is nothing that requires the inclusion of specific header records, or the names that should be used, and it is difficult to associate records providing details about a parameter. Also, variation in the names used for a parameter can cause problems to analysis software even though they are really just synonyms. There are similar in other file types unless standards are followed.

In a relatively new text-based format recently developed by the IVOA – VOTable – in addition to the name, type and units of the parameter, the FIELD records contain UCD and *utype* parameters. Using these – the *utype* is derived from a data model – it is possible to unambiguously define what a parameter means and it if done correctly it should be much easier for an external person to pick up and use the data.

Is it time for a new file format?

It may be appropriate to ask whether it is time to consider developing a family of new file formats that are better suited to interoperability. Existing formats – FITS, CDF, netCDF, Text, etc. – have their merits and are suited to certain types of data but the formats are now decades old and were not created with interoperability in mind. It is frequently difficult to just open up a file and know what it means and what it contains particularly since not all formats require that the parameters are properly annotate (unambiguously describe).

There is a tidal wave of data starting to overtake us and we must maximize interoperability before we are swamped. While we are not suggesting that all providers should switch to new file formats for existing data, they should consider this for new data that will in time overwhelm what has gone before. The virtual observatories should be able to manage to use of existing data sets on behalf of the users.

Data Provenance

If there are multiple copies of the data, it is essential that there is clear understanding of which copy is the master and which are slaves. The master should always have the most recent and most complete copy of a given dataset. The system should then provide the means to track which data are available on other sites and which version of processing they represent.

This issue has not been handled properly (if at all) in the past and should be addressed for future datasets. Most archive "mirrors" have been constructed without any consideration of the necessity to relate the holdings at one site with those at another. In most cases it will be necessary to generate provenance information from the files that are stored before such a comparison can be made.

Note that the interpretation of the meaning of provenance seems to differ between the particle physics community and our domain. The European Data Grid was designed around the needs of the LHC at CERN and in that context they are used to extracting clumps of data for analysis that they wish to keep track of; in our context we are talking about entire archives. Unfortunately a lot of the work that has been done in this area has been on behalf of the particle physics community and it is flavoured by their needs.

Resource Management

Resource Description and Registry

Registries provide a mechanism through which VO applications can discover and select resources. The Registry describes the resources that are available and how to use them – i.e. the nature of their interfaces.

The IVOA has developed standards to describe resources – these can be found under the *ReR* section on the IVOA Documentation page (http://www.ivoa.net/documents/). It also has a design and interface specification for the registry based on work done under Astrogrid.

Exactly what should be in the registry is not clear – different projects are interpreting the need for a registry in different ways and are placing different types of information in them. Even within a project, opinions of the type of information that should be included differ – how much detail should be in the registry and at what point should the individual service be consulted.

While this might at first glance appear to be an issue that only relates to the projects individually, if the desire is to share resources between projects across domain boundaries then having a standard set of metadata that can be exchanged is a major issue.

In discussion with the IVOA it is recommended that *the contents of the registry should be structured so that they require minimal changes* and that, where necessary, additional information should be provided by the service Support Interfaces (VOSI[11]). This simplifies the maintenance of the registry but implies that the type of information contained in a database should be described in the registry but not the details of the contents; this should be provided through the VOSI.

[It should be noted that HELIO has encountered difficulties in describing some of its services based on the types defined in the IVOA standards – these are mainly for services that perform some type of (processing) action to create a product. It has also been necessary to make an extension to better handle the multiple instances of each service that HELIO provides. How to

[11] IVOA Support Interfaces (VOSI) – http://www.ivoa.net/documents/VOSI/

describe the many, diverse tables in services like the HEC is also an issue – the VOSI interface may need further refinement.]

Authentication and Authorization

The purpose of Authentication and Authorization is to maintain user identity and control access to certain resource.

In order to ensure that users are not discouraged from using a research infrastructure, it should be possible to use most capabilities without needing to authenticate or even provide an identity. If a user wishes only to return to pick up sets of results that were generated earlier then a temporary identity (possibly supported by cookies) should be sufficient. However, if the user wishes to store data or user preferences on a semi-permanent basis, or use processing capabilities, then the system needs to know who they are – i.e. their identity.

For a few activities the user also needs to be authorized; these are limited to executing user-defined code (which could endanger the system) and storing user-defined material (which could have inappropriate content).

The metadata used to describe Authentication and Authorization needs to be domain independent and this should be relatively easy to achieve. There are a limited number of well established techniques for this and a project would normally choose to adopt one of them rather than trying to develop its own; this is very necessary since the effort required to get this type of token accepted worldwide is substantial.

Communities this document relates to

Heliophysics Communities

This document builds on the work of the HELIO, Europlanet RI and SOTERIA project. It is therefore mainly targeted at the communities that constitute Heliophysics – that is, the solar, heliospheric, planetary and geophysical communities. For the last two, it is mainly the magnetospheric and ionospheric observations for planets with a magnetic field and/or an atmosphere.

Solar phenomena that result in emissions that propagate through the heliosphere cause effects that could be observed in one of more of the domains. There is therefor a strong need for data and services from these communities to be interoperable.

Related Communities

The most obvious related community that CASSIS needs to interact with is the *geoscience community*. Emissions from the Sun affect the Earth in various ways and these could be explored if we could improve the interoperability with the geoscience community.

This community is working to an ISO standard (191xx) that facilitates the interoperability of the Geographic Information System (GIS); this potentially means that interoperability of heliophysics with large parts of the geosciences could be facilitated if we can understand how to work with this single standard.

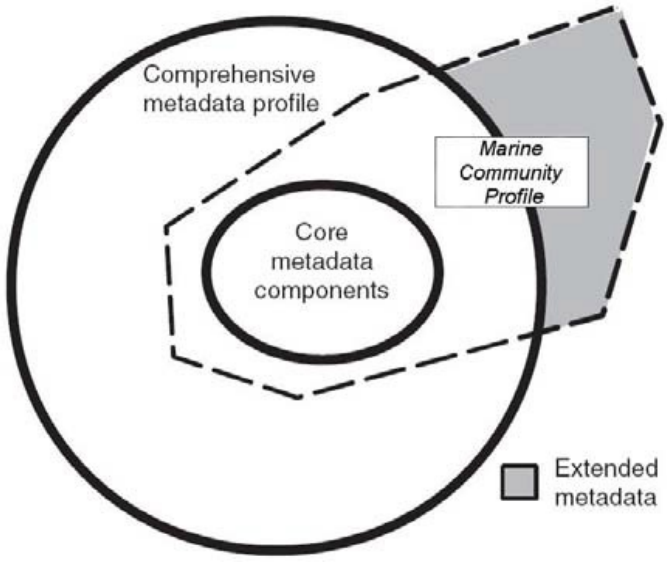

Figure 1 Illustration of how the profile created by a sub-community within the geoscience community fits within the overall scheme of ISO 191.

The ISO standard allows the community to make adaptions to meet their needs. Below is a short overview describing how the **Australian Ocean Data Centre Joint Facility** has defined a marine community of practice *metadata profile* of the ISO 19115 standard to

support the documentation and discovery of marine datasets – the figure and text are taken from *Marine Community Metadata Profile of ISO 19115*[12].

> The International Standard ISO 19115 *Geographic information – Metadata* defines around 300 metadata elements, with most of these being listed as optional. The ISO Standard specifies the process where individual communities can develop a "community profile" of the international standard. A community can adopt parts of the standard and also extend the elements, keywords and code tables to suit that community. The Marine Community Profile is compliant with ISO 19106 *Geographic information – Profiles* that describes the rules for developing profiles of the 19100 series standards.
>
> The Marine Community Profile is a subset of the standard and includes all ISO 19115 core and mandatory metadata elements. In addition, the Marine Community Profile has defined supplementary elements, code lists and controlled vocabularies to assist in the description of marine resources. The diagram illustrates the relationship between the core metadata components, the comprehensive metadata profile and the Marine Community Profile (*adapted from ISO 19115:2003*).

The way that a profile can be used to tailor which parts of ISO 191 are used by a community is analogous to the idea of having a common area and a domain or instrument specific area in the file metadata that CASSIS is proposing.

[12] This document can be found at www.aodc.gov.au/files/MarineProfileInfo.pdf.

Appendices

A) List of HELIO Services

HELIO Services are mentioned throughout this document. The table below shows the current list of services. An up to date version van be found on the HELIO Web pages[13].

Service Name	Purpose
Search Metadata	
Heliophysics Event Catalogue (HEC)	Maintain and provide access to *existing* event data from all domains
Heliophysics Feature Catalogue (HFC)	Maintain and provide access to existing feature data from all domains
Data Evaluation Service (DES)	Allows the user to create an auxiliary event list based on a *newly-derived* parameter, etc.
Context Service (CXS)	Provide context information to help the user make a selection
Review suitable observations	
Instrument Capabilities Service (ICS)	Match required observation type to one or more instruments (each part of an observatory)
Instrument Location Service (ILS)	Determine the location of an instrument (part of an observatory) at a specified time
Unified Observing Catalogue (UOC)	Provides information on whether specific instrument was making suitable observations at a specified time
Locate and Retrieve the Data	
Data Provider Access Service (DPAS)	Provide integrated access to data archives in all domains no matter how the data are stored or accessed
Enabling Services	
HELIO Registry Service (HRS)	Maintain and provide access to a registry that describes all the services available to HELIO
Community Interaction Service (CIS)	Manages interactions with the community, including authentication and usage statistics
Processing Service (HPS)	Support processing on demand
Storage Service (HSS)	Provide storage for user information
Coordinate Transformation Service (CTS)	Translated between the different coordinate systems used by the communities
Semantic Mapping Service (SMS)	Maps terms used in the metadata from the different communities
HELIO Monitoring Service (HMS)	Keeps track of the status and performance of the services that the HRS knows about

[13] HELIO Services – http://www.helio-vo.eu/capabilities/system-services.php

B) File Metadata

Parameter Naming

The names of parameters should be meaningful and should be used consistently throughout a data system – this should be defined in the data model.

In order to ensure interoperability, names in the common part of file should conform to a standard data model.

Some file formats have built-in complications in regard to naming.

Astronomically images in general, including solar images, are traditionally stored and distributed as FITS files. In its simplest version, a FITS file consists of a human-readable ASCII header of keyword-value pairs containing the metadata, followed by a binary part containing the actual data; a FITS file may also contain several extensions, and each of these may contain a data object.

Restrictions in the FITS format on the length of keywords limits the ability to create meaningful parameter names – this has resulted in some unintended difficulties. For example, there are many different contractions of the word wavelength in order to make it fit into 8 characters; there are additional difficulties if the name of the parameter is also being used to express a relationship with other parameters

Several attempts have been made to standardise the keywords, ranging from full FITS standard keywords, IAU approved keywords to more discipline specific keyword dictionaries. In addition, specific papers exist for describing world & celestial, spectral and time coordinates in FITS (http://fits.gsfc.nasa.gov/fits_wcs.html).

This demonstrates an inherent weakness of the FITS format – it is too loose a standard with few rules to constrain it formulation. There is no formalized set of names for FITS keywords (although there are many recommendations). Also, nothing defines the relationship between the parameters, if there is one. An external piece of information is required to decode the file – this is assumed knowledge.

Parameter Annotation

Metadata are more interoperable if the parameters are fully and properly described. The VOTable format developed by the IVOA comes to our aid in this context; following this structure could be beneficial to the objective of interoperability.

The VOTable format is an XML standard for the interchange of data represented as a set of tables. In this context, a table is an unordered set of rows, each of a uniform structure, as specified in the table description (the table metadata). Each row in a table is a sequence of table cells, and each of these contains either a primitive data type, or an array of such primitives.

The table metadata is the header section of a VOTable file that contains a field record for each parameter. The field record can have many components: the most obvious are the name and a description; other components can include the units and two that are known as a UCD and *utype*.

The Unified Content Descriptor (UCD) is a formal vocabulary for astronomical data that is controlled by the International Virtual Observatory Alliance (IVOA). The vocabulary is restricted in order to avoid proliferation of terms and synonyms, and controlled in order to avoid ambiguities as far as possible. It is intended to be flexible, so that it is understandable to

both humans and computers. UCDs describe astronomical quantities, and they are built by combining words from the controlled vocabulary.

In many contexts, it is important to specify that parameters convey the values defined in an external data model; the *utype* attribute makes it possible to unambiguously define the meaning of the parameter. The *utype* attribute is especially useful to specify the spatial and temporal coordinates present in the table when it contains astronomical events: these parameters are essential to most applications that process multi-wavelength data.

External File Descriptor

One solution to the difficulties discussed above is to define an external XML file that provides a standard way of describing the metadata and maps into the file headers. This would be associated with the data file by including a link in the file header; potentially a "style" file needs to be associated with each dataset.

Such a solution would be a way of increasing interoperability and should be able to accommodate any type of file format. It could also allow older data to be made more equivalent to more recent data.

The XML structure of the file would make it possible to group related parameters; the names used in the data file as well as their standard equivalent names, UCDs and *utypes*, etc. could be included and hence the meaning of a parameter and it relation to a data model would be fully described.

This approach would make it easier to create a standard tool for handling all types of files – information of how to translate the header to standard terms is provided by the XML file.

While it would be beneficial to re-examine the file formats that are being used and possibly introduce new formats that are better suited to interoperability, the approach could reduce many of the difficulties we currently face.

This concept is now being explored within several projects and will be tested using data from the SWAP and LYRA instruments on PROBA2.

Parameters that should be included

A list of recommended parameters is being prepared and will be included in this document if it is re-issued.

C) Rules for data storage

Assuming that using data across domain boundaries can be made easier by improving the quality of the Observational Metadata, the next task is to make them more accessible.

Following a few simple rules on the way files are named and in the directory structure used to hold them can make a significant difference to how easily they can be accessed.

There are no hard and fast rules for file names, but the names should be sufficiently unique that files can be stored outside of their "native environment" – i.e. where they are normally stored on the system of the group that generated the file. In other words, the file should be able to exist without causing confusion when removed from the context of where it is normally stored .

The file names should also ideally identify the type and origin of the file, the nature of the data and when the observations were made; the name might also indicate if an image represented a partial or the full field-of-view of the instrument.

Although not strictly metadata in itself, the structure of the archive can make a lot of difference to how easily the data can be accessed. If the data are held in a hierarchical structure based on date and time, it is much simpler to create metadata that can be used to find the required data. Storage with this type of structure is essential for resource-poor providers and would be beneficial for data centres.

A summary of observations that have been made, or Observational Catalogue, is useful and simplifies access. It is particularly useful if not all the observations in the archive are available on-line.