

# Lessons Learned from Solar & Heliophysics

## Abstract

This is a collection of observations and insight from six years of working on the Virtual Solar Observatory project, interactions with members of the Heliophysics Data and Modeling Consortium, and other earth and space science informatics efforts.

## About the Author

I entered into this field purely by accident—I was hired to write the real-time data system for the Solar Terrestrial Relations Observatory (STEREO), but one of the programmers on the Virtual Solar Observatory (VSO) handed in his resignation on my second day, and I was quickly reassigned. The VSO is a distributed federated search system for solar physics data [Hill, et.al, 2009]. As such, I have written a number of translation layers so the VSO can talk to the different data catalogs maintained by solar physics archives, and have seen some of the differences in how scientists organize their data. I am responsible for many aspects of the search system from interfaces to the data catalog for the STEREO Science Center. I also participate as the VSO's representative to the Space Physics Archive Search and Extract (SPASE), an effort to create standards for heliophysics archives to provide descriptions of their holdings and allow searching through a single protocol.

When I started, I had very little experience in dealing with scientists or their data. Although it gave me an advantage in that I wasn't tied to the same pattern of thinking as those already familiar with the discipline's data, I made a lot of mistakes in my assumptions about how to deal with the scientists and their data. Hopefully, this article will help others to avoid some of the errors that I made.

I'm writing this in the first person, as I want it to be clear that these are personal observations and this is not the product of an ethnographic study or other research effort. It might be possible that my experiences are completely abnormal and not applicable to the larger scientific community, but if that's the case, I hope that you can at least find some amusement from my anecdotes and be thankful you haven't had to deal with some of the situations that I've suffered through. I'd love to have a coherent research statement and a nice, neat conclusion but most of my time is focused on implementing data systems, rather than abstract study of the field.

## **Understanding the scientists**

Before we can even consider dealing with data, we must understand the people that we're going to be talking to as we attempt to document their data. First, we have to be careful about the language that we use, as there are a number of terms that might have subtle differences in meaning than what we are used to; I specifically said 'document' the data and not 'catalog', as 'catalog' can be a specific type of scientific product that are the result of scientific research and are frequently peer-reviewed. Although we can claim to be creating a 'data catalog', to

differentiate from a 'feature catalog', 'event catalog' or other 'science catalog', it's safer to just avoid the term when talking about non-scientists creating indexes for the data.

We must then prepare ourselves for some potentially strange personalities. As the scientific method involves trying to disprove a null-hypothesis, you should be prepared for some of the scientists trying to find flaws in anything you say. Even if your intent is worthwhile, some will disagree with your implementation, and by correlation, argue that the effort overall isn't worthwhile. When dealing with the Primary Investigators (PI team), there may be one scientist who is considered to be the local expert or self-appointed czar in some aspect of information technology, be it hardware, databases, programming or whatever; if such a person exists, you may get push-back when talking about data systems as they attempt to assert their dominance in the topic.

With 'big science' efforts, the science team may have been working together for years before they begin collecting data. If we just look at the PI team, they wrote their proposal, went through the award process, built their instruments, tested the instruments and finally installed and calibrated the instruments. If there's a spacecraft or a fixed observatory involved, there are additional delays for launch and site construction. These efforts take years before any data is collected. For the Solar Dynamics Observatory (SDO) which launched in February of 2010, the instrument teams submitted their proposals in 2002.

This means that by the time that when I've started working with the groups to find how they're going to be handling their data, I'm viewed as an outsider, and have to gain their trust and respect. I'm not sure if it was a problem with my lack of presentations skills or if scientists just like being argumentative, but with some groups, I've had multiple hours of discussions on what I thought would be relatively straight forward modifications to data systems in which the discussions took more time to discuss than to implement. Even for code that I'm writing, I have spent hours explaining what I'm planning on doing, and arguing if I'm doing things correctly.

This makes sense, as scientists are expected to present and defend their research and findings, but I find it can be emotionally draining. As you gain the respect of the scientists, you might get more leeway and freedom to implement something so you can at least give them a prototype for them to criticize, rather than have them assume where you're going to make mistakes, but for some groups this might take a year or more.

The easiest scientists to deal with are those that come to you—if you can establish your reputation, and become accepted as an expert in the field, people will come to you for your opinion. In my particular case, I was actually riding off of the reputation of the established scientists on our team, and we integrated our own data holdings before we sought out other organization's data to add to our federated search. As we grew, some archives saw the benefit of joining and came to us—and it's much easier to work with a willing participant than a group who is unwilling to accept that other scientists might want to interact with data in different ways than their data system was designed to support.

## Understanding the data systems

Just as the data is heterogeneous, so are the data systems [Hourclé, 2008c]. This is accepted because the needs of every science discipline has different needs: “Because of the variable complexity of solar-terrestrial research problems, no single data management concept will dominate, and features of many approaches will be required in varying proportions for specific situations.” [NRC, 1984, 3]

Some data archives are simply an FTP site and a few directories of files with no dedicated catalog. Others might have a database or simply ASCII tables to catalog the files they distribute. Yet other data systems only generate files on distribution, and keep all of their data in a database so it can be subsetted, processed and packaged on demand. Data systems for newer data systems track ingest more observations in a day than others have over years of operation.

Compared to the effort that is put into building and testing the sensors, the data systems are a relative afterthought. It is common for the data systems to be slight improvements over previous data systems, or to take the data system designed for one instrument, and attempt to shove the data from one instrument into a system designed for a significantly different type of instrument. In the case of SOHO, the mission database was designed for what was believed to be the most complex instrument to describe, and then forced all other data descriptions into the database. To achieve this, each database field means slightly different things for each of the twelve instruments.

This isn't a new issue; it was identified in 1982 :

“There is a commonly a lack of scientific involvement in data-system planning during early mission planning and during the system development phase. Typically, the interdisciplinary nature of data is not fully recognized, and, therefore, data systems are frequently not implemented for their actual use.” [NRC, 1982, 2]

I would say that the current problem is that there is too much scientific involvement [Hourcle, 2009b]. This involvement tends to be shallow without interdisciplinary input, either by other scientific disciplines or with experience in data modeling or other information fields. Most PI teams tend to be composed of experts in their field and are less likely to consider that other scientists will attempt to search and use their data in a different manner than the main investigation. In the case of the SDO AIA instrument, because the data was being inserted into a system originally designed for SDO HMI, in the preparations to served data to the general community, VSO team members realized that the original plan for the data's organization would require that a scientist requesting data to be generated in a single wavelength would place eight times the load on the system.

Because the PI team planned to pack the images into ten second blocks, and there were eight images in different filters for every ten second period. The instrument team didn't see this as a problem, arguing that other scientists shouldn't be looking at a single wavelength, but should be looking at events in all filters. This may be the case for in-depth solar physics analysis, but as the sun is a driving force into all other systems that are studied in earth and space physics, the

data is desired by a great number of scientists who may not be attempting to do the same type of analysis that the instrument was planned for. Other scientists may be interested in browsing using a limited view of the sun, and then expanding their search after finding a period of interest:

“Usually, data archives do not include an adequate browse capability. Such a facility would allow the interested user, at his home institution, to locate and inspect data sets rapidly and to select those that will be useful for further analysis.” [CODMAC, 1982, 4]

The PI built data systems have well known problems:

“The PI, usually a scientist, has generally been involved from the beginning through the end of a mission. At times, however, problems arise because too little thought resources have been given to data management in the planning of the mission. ... The main task for the PI has been to reduce the data for his own needs. As a consequence, the data are sometimes not useful to, or not interpretable by, the rest of the scientific community.” [CODMAC, 1982, 140]

These statements might be almost thirty years old, but the problems still remain due to a lack of proper attention given to data management [Norris, et.al, 2006].

## **Understanding the Term ‘Metadata’**

As we’re talking about e-Science metadata, I feel it is important to define what I qualify as metadata as one person’s metadata is another person’s data. For the communities that I deal with, the ‘data’ are the processed or unprocessed values from sensor recordings, while the metadata are the information needed to understand the values. The metadata might be stored as headers within the data file, in an ancillary file, or in software used to process the data.

For solar images, scientific metadata might include the camera's location and pointing, the time of the observation, any filters or polarization that might affect the light entering the sensor, temperature and other issues that might affect the sensitivity of the sensor, details about the processing and handling of the data, or information about how the data is packed within the file.

Although we could attempt to define metadata based on the sections within standard scientific file formats, there may be headers and sub-headers, or multi-dimensional metadata packed within the payload of the data, such as image masks or maps of the error within an image. The boundary between data and metadata gets fuzzier the more highly processed the data is, but from my experience, the scientists expect 'science data' to be directly derived from the original sensor recordings; it might be reduced or transformed in some way, but it comes from the sensor. Everything else that isn't the data is therefore the metadata.

## **Understanding the 'Archives'**

Science archives may not fit the classic definition of 'archive' as used by the museum and library communities. Some archives are 'active archives' in which the data is being collected for an active science investigation. As scientists learn more about their instruments and the field they are investigating, they may recalibrate the data already in the archive. In some cases, the files given a particular identifier may be modified; in others, the data archived remains the same while the metadata is modified.

This often results in older versions of the data being removed; due to the high volumes of data and the lack of perceived value of the knowingly incorrect values, there is little reason to maintain the data. Some systems will include the calibration version in the identifier, so we might realize that the previously obtained version is missing, and find the most recent version. Unfortunately, it is also likely that identifiers based on time might change, making it difficult to verify which is the correct replacement for the previous data.

For solar images, as the telescopes will degrade over time, possibly in ways that are not caught by the standard calibration process. In solar physics, scientists often prepare for this possibility, and thus store the raw images, while maintaining separate catalogs of the correct metadata. Unfortunately, those catalogs may be stored in proprietary file formats that require specific software to access [Hourclé et.al, 2007].

In other cases, as new processed forms of an instruments data are available, or as the data volume reaches limits of the storage system, the data might be rearranged; as many systems rely on file paths for file identification, this can result in an apparent deletion and creation of new data.

There do exist 'final archives' or 'deep archives' that act as a data mausoleum, with the primary focus being on long-term preservation rather than use of the data. In some cases, the data provided to the final archive might not have the proper documentation to be usable, and are simply stored until a data recovery effort might occur.

## Understanding the scope of the systems

You would think that for a ‘Virtual Solar Observatory’, we’d at least have a vague agreement of what a ‘solar observation’ is. Unfortunately, the issue is much more complex than that, as there are a number of observations of solar phenomena that aren’t specifically observations of the sun. The original scope of the Virtual Solar Observatory was that we served ‘science quality’ solar physics data. This has made for a rather poor definition of our scope as ‘science quality’ means something different for scientists trying to understand the inner workings of the sun as it does for those trying to use observations of the sun to predict events that might affect the earth (aka. ‘space weather’) or for those who study the sun’s effects on the earth and the rest of the solar system (aka. heliophysics). All have different requirements for quality, and what is useful for one discipline may not be useful to another.

Even if we were to take a very minimal view of ‘solar physics’, the needs of scientists planning instrument campaigns more closely aligns with the needs of the space weather community, where the age of the data is more important than the absolute calibration and amount of error within the data. As our scope changes, the metadata needs change, as each community has different aspects of the data that they are concerned about. Sometimes, we simply need to differentiate between the community that the data serves.

My suggestion is that whatever metadata you decide to include in your schema, that you do not require data providers to supply any information other than that used to make their data findable

and usable to their local community. Although we do want to make the data useful to other disciplines, setting too many requirements will create extra burden on the scientists who may not even understand the full implication of the metadata used by other disciplines. You will occasionally find scientists who specifically do not want their data to be used by other disciplines, as they've had previous experience with people misinterpreting their data; although this should be an argument for better documentation of the data, you are better off getting what compliance you can from them, and have others attempt to fill in the missing metadata later.

Luckily, most scientists can at least agree on what an observation is. Unfortunately, that's not that most data systems are designed to track.

## **Understanding the data system's records**

Just as with library catalogs [IFLA, 1998], each scientific data system can be tracking slightly different definitions of 'data' [Hourcle, 2008b]. Most of the data systems I have run into track files, not observations. As there is a many-to-many relationship between observations and packaged files for distribution, we often have problems trying to de-duplicate the records to select individual observations, particularly across archives [Hourclé, 2008a].

As there is no standardization in the language used to describe the objects being tracked within a catalog, trying to discuss data catalogs can cause additional confusion [Hourcle, 2008c]. The term 'data set' is used in the space sciences to refer to a collection of similarly data from a given experiment, while in the earth sciences it is a single object, with the collection being a 'data

product' and the discrete objects to be distributed being a 'data set'. A 'data product' in solar physics is a discrete object. This creates problems such as meeting on virtual observatories in 2005 where I agreed with another person that we needed collection level metadata registries, and that we wouldn't be able to easily describe each individual object in our archives [NASA, 2005, 15-16]. Unfortunately, as we used the terms 'data set' and 'data product', we vehemently agreed with each other, wasting almost half a day of our session on metadata registries.

In some disciplines the terms are considered synonyms. My advice is to avoid the terms 'data set' and 'data product' entirely. 'Data collection' or 'data series' are less ambiguous, and either 'data file' or 'data object' are more clear. You will occasionally encounter the term 'data granule', which I've had defined as the 'smallest amount of useful data' but as data can be used my more than one discipline, 'smallest amount' is entirely arbitrary. Others define 'data granule' it as the 'smallest value separately addressable' but the more accurate definition might be 'the smallest amount that we bothered to track in our system'. Depending on the field, a 'data granule' might be composed of multiple 'data records', which for the most part correspond to the individual observations. We also tend to avoid the term 'data granule' in solar physics, as a 'granule' is a type of solar feature.

Other fields may track their data by discrete files, but they may track multiple observations per file. For heliophysics time series data, the concept of 'observation' isn't typically used; as one scientist explained, "Data are acquired on an ongoing basis and may be 'decomposed' into segments by time or location of observation, but virtually never by 'observation'" [King, 2007].

## **A comment on standards**

There are a lot of standards for scientific data and metadata ... dozens of standards for packaging, transmitting and querying. [Hourcle, 2009a] Part of the problem is that different disciplines think about the world differently, and have different needs for documenting, searching for and using their data. [NASA, 2005, 15] What might be optional metadata in one domain might be required for understanding the data in another, and each discipline defines their standards to create the most benefit for their field without requiring an unreasonable burden on their members.

It is unlikely that we will ever get to a single standard, unless it is defined in such a way as to make it extensible to the point that there are multiple variants that allow each discipline to customize it to their needs. Of course, this defeats much of the point of standardization, as we still have to deal with interoperability between all of the different variations that might arise.

Attempting to make general standards that provide a general level of understanding of the data set is one way to avoid disagreement about the standard, but something that is not specific enough for scientists to find useful will likely not be adopted without being forced on them by funding agencies. Even then, the scientists will complain as they have to maintain multiple descriptions of their data and drag their feet in complying with the mandate if they don't see a benefit to it.

## **Too much metadata?**

As much as we'd like for all possible information about the data to be recorded, there can also be problems with too much metadata within the data files; although the level 0 SOHO/EIT data served by NRL is more complete as it contains engineering metadata needed by the PI team, some visualization tools crash as they can't handle the metadata.

In other cases, the tools just ignore metadata that they don't understand. Sometimes, that metadata is critically important to properly understand the data. In the early days of the STEREO mission, the STEREO-Behind spacecraft was flipped over so that its antenna was pointed towards Earth. Unfortunately, the generation of JPEG images for browsing by the mission archive didn't properly handle the field, and the JPEG images were generated with solar north at the bottom of the image. As each of the individual images looked okay, the problem wasn't noticed until the images were viewed as a slideshow, and the sun appeared to rotate in the wrong direction. This one piece of metadata could have resulted in dramatically different interpretations of the data and incorrect scientific conclusion.

## **Documenting the data**

As the terms used to describe the data objects vary so significantly, we will focus on three major objects to be described—the observation, the scientific data, and the packaging of the data. This is similar to my previous attempts to propose an alignment of science data cataloging with FRBR [Hourclé, 2008b], but there are some issues with time series data and other collections and

aggregate objects that still need to be resolved to make it more universally applicable. For this paper, the three objects described most closely align with my earlier definitions of Observation, Expression and Manifestation.

(figure1)

FRBR Applied to Scientific Data [Hourclé, 2008b] (updated for this article)

Observations are effectively a type of event, and so we can ask the basic journalism questions of who, what, when, where, why and how. The scientific data are the processed values used to express the observation, and thus we need to describe the processing that has been applied, the units that the values are in, and the known error. The packaging can affect how easy it is to use the data, as the file format affects how easily a scientist can use the data with their existing tools and how easily it can be to obtain.

removed : Some metadata standards will group their metadata fields into general categories, such as identification, quality, distribution, time, citation, etc. [FGDC, 1998]; while others might break it down by user, such as engineering vs. science metadata; or by when the metadata are used, such as finding vs. use of the science data. The standard might simply define the groupings by what they don't include—the SPASE standard includes 'use caveats' for finding data that is useful to a researcher, but there was a group decision to leave other aspects of use, such as structural metadata, to other standards.

### **Describing the Observation**

Just as with journalism, the basic questions of who, what, when, where, why and how are useful for identifying data. Who generated the data, what was being observed, when did they observe it, where did they observe, and how the sensor was observing are universally useful for describing data, although some disciplines might have assumptions that keep them from always listing this information. Knowing why the data was collected is useful for setting context and

defining the use caveats, it doesn't tend to be a major discriminator in determining if the data is of interest.

### **Who?**

Although the PI team are important in 'who' is observing, the more important item to describe is the detector that actually generates the data. Unfortunately, there are a number of overlapping terms and concepts used to name the detectors: instrument, telescope, detector, camera, and sensor. I will not attempt to define any of these terms, as they are used in a number of conflicting ways; I've even seen the spacecraft name recorded in the FITS 'telescope' field. The name given to the hardware might match that given to the investigation or experiment, but it doesn't need to. There are also times when the names of the mission and associated hardware will change over time; most commonly before launch, such as with RHESSI (aka. HESSI), or immediately after launch, as with all Japanese spacecraft, but there are cases where spacecraft have been renamed after data collection has started, as with Fermi (formerly GLAST), and times when the mission name changes as a spacecraft completes its original mission and is repurposed.

Some instruments are composed of more than one telescope, camera or detector, while in other missions each individual detector is considered a separate instrument. These component parts of an instrument might all be similar and just have different spectral sensitivity or they could each collect different types of data.

In some cases, the name of the observatory or spacecraft is enough to describe the detector, as they might only have a single telescope; but for ground-based observatories with only a single telescope, the observatory name might not be enough, as the telescope's optics or camera might be changed in the future. Occasionally, a camera will be moved to a new observatory, but we can normally consider it to be a new camera, as it will need to be recalibrated after it is relocated. In other cases, the detector name isn't sufficient to uniquely identify the detector, as there may be multiple detectors built and installed in different locations, such as with the STEREO and Cluster missions or the Precision Solar Photometric Telescope.

### **What?**

One issue with dealing with multiple disciplines is how they think about what is being observed. There are a wide variety of types of detectors, and some are *in situ* while others are remote sensing. For the disciplines that deal with mostly *in situ* observations, they tend to search in terms of the location of the sensor, although same as the location being observed, whereas remote sensing observations are more concerned about if a given location was observed and then where it was observed from.

The issue of 'what' is being observed is that there are a few different concepts that we need to track; for some observations, a sensor might continuously observe a region and wait for something to occur. In this case, we can easily define the region, eg, 'the eastern limb of the sun'; we might also have the campaign information so we know what features or events the

instrument team was attempting to observe, but defining features that were seen by the observation is a more ambiguous process.

In some cases the PI team will record what they believed they saw in their file headers, but in other cases, the identification of the features and events are stored in separate catalogs. These catalogs are the product of scientific research, and the identification process and resultant catalogs are frequently published in peer-reviewed journals. Because of this, it is important for library catalogers to not use the term 'cataloging' when discussing the process of describing science data; some scientists will take much offense if they think that a non-scientist is going to be making judgments about what features or events are being observed within the data.

As the data is re-calibrated, it is possible that additional features will be recognized by the scientists, or that formerly identified items will be removed as incorrect; with the large data volumes of some instruments, there are data pipelines that will attempt to do image analysis in software, but scientists will review the results to remove false positives. For some types of features, there will be multiple science teams using different identification methods; they might use the same or different data, they might process it differently and they might use different qualifications for their feature identification, all of which results in disparity in the identification of phenomena.

**When?**

Another universally needed parameter to make sense of data is the time. Time includes both the duration of the observation (exposure time), and the date and time that the observation was taken (observation time). Unfortunately, in practice the 'observation time' might either be the beginning of the observation or the middle of the observation; you might find the temporal information recorded as the start time and duration, the start and end times, or the duration and midpoint of the observation. If the exposure time is fixed for the instrument it might not be recorded with the discrete observations, but with other documentation for the instrument as a whole.

The precision needed for temporal information varies greatly between disciplines. For observations of relatively fixed objects, scientists might not need the exact second that an observation was taken. If we are attempting to track fast-moving objects, or looking at a changing object the need for precise timing becomes more important; milliseconds or even nanosecond precision might be needed.

If we are attempting to coordinate observations between multiple sensors, we also need to ensure that the times are accurate to an external reference, which creates a new set of problems. Time is frequently used as the primary identifier for observations, but times might be adjusted as part of the calibration of the data; as no clock is truly precise and spacecraft may be undergoing relativistic effects, the spacecraft time will slowly drift away atomic clocks on earth. As these issues are corrected.

You may occasionally encounter time zones and daylight savings issues, but to simplify coordinate with other instruments, many will use Universal Time (UT). Unfortunately, there are a few different UT times, as there is Coordinated Universal Time (UTC), which is based on time from atomic clocks but has leap seconds inserted to keep it close to UT1, which is based on the Earth's rotation. Instruments are most likely to count some number of seconds since an epoch, without consideration for leap seconds. With this, we can easily convert to International Atomic Time (TAI) or Global Positioning System (GPS) time, but we must maintain a table of leap seconds to compare the time to UTC, which is the reference frame that some disciplines expect.

Time can become even more confusing when coordinating remote sensing of distant objects, as with the STEREO mission. Because the two spacecraft are not an equal distance from the sun, to capture an image of the the same state of the sun, the spacecraft take their images a few seconds apart. To compare STEREO observations with other spacecraft or ground based observatories, these issues need to be taken into consideration. [Thompson and Wei, 2009]

### **Where?**

The problem of describing where we are observing comes down to coordinate systems -- we need to express the location of what is being observed, but there are different needs depending on the field. As most disciplines deal with a single coordinate system, it makes sense for them to set the coordinate system for their domain, rather than expecting everyone to convert as necessary.

There are standards that handle multiple coordinate projections, such as the World Coordinate System [Mink, 2002] but being able to specify the location does not make it easy to search on. For each projection, you might have different reference points; for something as simple as the two dimensions of a digital image, we could declare any one of the corners, or even the center of the image to be our zero point.

Within the VSO, we have discussed coordinate systems for years and our problem has been that there are multiple coordinate systems typically used in solar images [Thompson, 2000]; most instruments are pointed in terms of x and y coordinates from the instrument's point of view (heliocentric-cartesian), but features and events are typically tracked in terms of Carrington longitude and latitude (heliographic) [Leibacher, 2009], to deal with the rotation of the sun relative to the observer and to make it easier to compare across instruments, or in terms of polar coordinates (helioprojective-radial) to deal with phenomena that are ejecting from the chromosphere.

Part of the problem is that Carrington is not just a spacial coordinate system, but is a spacial-temporal projection. This means that the conversion between Carrington and the other coordinate systems depends on the time of the image. For search queries, this means that a search in terms of Carrington coordinates for an extended time period will create a sliding window in other projections. Although we could attempt to work in a single coordinate system, each conversion reduces the accuracy of our searches. Although it's worthwhile in some cases to record the observation in more than one reference frame, it is unrealistic to expect us to enable to

searching in the potentially hundreds of different coordinate systems, with the majority of them making no sense for a given observation.

### **How?**

Some sensors have multiple ‘observing modes’, and which mode is being used affects how a given instrument observes its environment. This could be as simple as varying the exposure times of the observation, the filter, or the polarization of the light entering a telescope. To speed up the time between observations, sometimes only part of the CCD is read out; in other situations, the CCD isn’t cleared out between readings, so to get the discrete observation, we take the difference of the reading the one before it. In some cases, such as with Solar Optical Telescope (SOT) on Hinode or the Helioseismic and Magnetic Imager (HMI) on SDO, there is more than one camera attached to a single telescope, and the cameras have dramatically different observing characteristics.

Based on the observing mode of the sensor and the characteristics of the sensor itself, we can get other critical metadata—what type of physical quantity it was measuring, what each dimension of the resulting data represents, what the spectral sensitivity was, etc. Of all of the groups of data mentioned so far, this is possibly the most difficult set of metadata to reach agreement on between disciplines, simply because how you describe a telescope is different from an antenna, magnetometer, spectrometer, particle detector, or any other type of sensor.

## **Why?**

Although it's not typically used by the generalist scientist for identifying observations, setting the context of the data collection can help to explain the assumptions made by the PI and determine caveats for interpretation of the data. As with 'how', the 'why' is typically a function of the primary investigation, but for those instruments that aren't running a fixed synoptic program, they might run a variety of observing campaigns. Unlike the 'how' aspect in the specific details of the cadence of the observations, the filters used, or the exposure times, the 'why' is external to the instrument itself.

Some missions have a 'guest investigator' program, where someone other than the PI team can propose an observing plan. Multiple instrument teams, possibly from more than one mission, may plan to observe the same region in 'coordinated campaigns' so that they can gain a deeper understanding. Although this information can be useful for finding observations by the guest observer, it has not seemed to be as useful for finding information, as the guest observer also knows what time period that had access to the instruments. The coordinated campaign information might be useful, but the information is typically stored as a free-form string or an identifier to a separate campaign catalog, and may not be useful on its own.

There is one exception, and that is if the campaign information can easily identify abnormal observing that might not be useful to the general public. Examples include 'darks' and other observations used to calibrate the instruments. If we can readily identify such observations, we can avoid sending them in response to queries from the general public.

## Describing the Scientific Data Values

How the recordings of the observation are processed into scientific data values affects its usability, and thus there might be multiple processed forms stored for different uses; but describing the processing and the resultant error varies by discipline.

deleted :  
Besides the earlier mentioned groups of data, there might be hundreds if not thousands of other bits of information needed to properly understand and make use of the data. Possibly the most important of these to the scientists is the amount of known error within the data, but there could be any number of assumptions made in the processing of the data that might affect the interpretation of the data.

### Error

For all scientists, the precision and accuracy of the data are critically important -- but how they express the error varies greatly. In tabular data, there might be an extra field to indicate the known error; in data plots, there might be error bars; for images we might have an extra layer within a data cube. Some fields rely on the error being handled by the data processing software, rather than being attached to the data.

Unfortunately, how each field expresses their error might not work for other fields. The FGDC standard specifies pointing error as being meters on the ground [FGDC, 1998]. This might work for remote sensing on solid bodies, but just doesn't work as well when dealing with astronomical images where each image includes a number of objects at varying distances. For telescopes, pointing precision is expressed in terms of arc-seconds; this works well for astronomy, where the telescopes are all in basically one place relative to the distance to the objects being observed.

In solar physics, as the raw data is distributed and the software to process it, there are cases where the data is known to have errors that can be corrected through processing [Metcalf, 2003]. Although a more correct version of the files will be generated for the final archive, this does not occur until after the mission has finished observing. As missions are extended, it could be a decade before the data is corrected.

### **Processing**

Although many of the scientific file standards define a metadata field to track the processing of the data, they are often free-text; as there are differences in the terms used to describe the processes. Some of the processes align with the relationships within FRBR [IFLA, 1998], but we may need more precise descriptions of the processes applied to determine if the data is of use for a particular type of research.

(figure 2)

Group 1 Entity Relationships [IFLA, 1998]. Format from [Hourclé, 2007]

Concepts such as ‘summarization’ of a time series of images could be time averages or pixel summing (aka ‘binning’). Attempting to identify what type of summarization has occurred would require analyzing the other metadata to determine if the cadence or pixel scale had changed from the original. Other process, such as ‘flat fielding’, ‘despiking’, ‘background subtraction’ and other methods to prepare the data for analysis simply don’t have counterparts in libraries. Some processes, such as ‘sampling’ (a form of data reduction in solar physics) is called ‘slicing’ (or

‘dicing’, if in multiple dimensions) in other fields where ‘sampling’ is the term used to describe data collection.

Not knowing the type of processing that had been applied to the data can result in incorrect science results or simply waste scientist’s time. In one case, a scientist told me that they had noticed when they compared calibrated images for a telescope, they noticed that the average intensity stayed the same over years worth of data; upon investigation, it was discovered that the data was normalized as part of the data processing, and meant that images from years apart could not be compared; the brightening of of the solar limb might be the result of dimming across the disk.

For some fields, this means that the raw observations are preferred over processed data, so that the scientists can be assured that they know what processing has been applied, and that it is applied consistently over all of the data being used in their research. For some instruments, the raw data is only useful by the PI team to generate higher level data products.

### **Processing Level**

There is some slight standardization in describing how processed the data are, but it is not directly comparable across missions or disciplines.

(table 1)

CODMAC Data Levels [NRC, 1982 via NRC, 1997, 16]

Some fields use 'level 0' to refer to the raw sensor data, with 'level 0.5' to denote what was sent down from the spacecraft if there was some sort of irreversible process applied. Some may denote a whole range of 'level 0' images (level 0.1, 0.3, 0.7), based on the metadata packaged with the data, while NASA's Earth Observing System defines this as 'level 1A' [NASA, 2010]. Unlike with earth observing systems, where observations can be compared to 'ground truth', solar telescope teams don't attempt to generate 'level 2' until after the mission is complete; level 0 (raw) or level 1 (calibrated) data are the typical norm.

A science team might also produce 'quicklook' data, which is calibrated, but in an attempt to produce the data as quickly as possible. In solar physics, quicklook data is used for space weather forecasting, but has not undergone the full processing and vetting, and so should not be used for in-depth analysis. The STEREO mission also has 'beacon' data which is near real time, but is highly compressed; the compression used results in a number of image artifacts which are highly sought after by people attempting to find signs of alien activity.

### **Describing the Packaging**

The packaging can be described in terms of the data format; many scientific data formats are 'self-describing', but this typically means that they have a mechanism for describing the data contained within them, not that they can necessarily be used without understanding the file format. Knowing the file format is typically sufficient to identify if the data will be usable by

their tools, although there may be a few different versions and variants of each file format; knowing the specific variant metadata profile is used within the file should provide the rest of the information needed.

To round out the information about the packaging, it is also helpful to provide the user with an indication of how the data are aggregated in the package, and the overall size of the data file; not all file systems or analysis tools can properly handle files over two gigabytes in size, even if the individual images within a data cube are within the limit.

## **Adoption of Standards**

As there are so many standards out there, the ones that will be most widely adopted are those that provide the most benefit, with the least cost to implement. Attempting to impose standards on groups will be ignored [NASA, 2009, 15], sometimes even with community input. As the needs of each community are different, the standard with the most benefit for one field might not be the best standard for another. If a community already has a working solution that does most of what they need, they have little reason to change to something new.

If you do have to create a new standard, keep it simple. Partly to gain support and adoption, but also “the imposition of overly ambitious comprehensive data systems can result in costly systems that do not address basic needs” [Sibeck and Kucera, 2002]. Much of the VSO’s success has come from keeping our standard simplistic compared to others, defining our scope (solar physics), acknowledging our weaknesses (the Virtual Heliophysics Observatory handles *in situ*

data better than us), and trying to solve as many of our community's problems without bulking up the standard by adding new elements that won't be used in most cases. Scope creep can be a vicious cycle—once you get up to a hundred or more elements, adding one more doesn't seem like such a big deal, even if most of the time we aren't going to need to track deceased people.

Make sure that your schema, vocabulary or other standard has a clear purpose and significantly improves over other alternatives in use. If something else suffices for their needs, they have little incentive to switch, especially if they need to add support for it in into their analysis and visualization tools.

If you are looking for opportunities, ask the scientists what it is that they can't do now, that they'd like to be able to, or that they can do, but the current effort involved just isn't worth it. If you can find a solution to one of those problems, you will gain support. Most scientists would rather be 'doing science' than trying to download data or reformat it for use in their tools.

## **Summary and Recommendations**

The science community recognizes the problems with so many different data standards and they are working on moving towards a reduced number, but it is highly unlikely that we will ever get to a single standard across all sciences for storage, querying, or any other task.

The easiest way to gain adoption of metadata standards is to provide an immediate benefit for compliance; building better tools for querying, visualization or analysis that make use of the

standard are as important as the standard itself. Reduce the cost of complying with the standard by keeping it simple to apply to data collections; make sure your documentation is clear, and provide tools and assistance.

Working with the scientists can be an interesting experience, but with enough time and patience, you can get some agreement on terms—at least until they come up with some novel experiment, try to work with another discipline, or just gain a better understanding of their field.

## Acknowledgments

Thanks to the scientists and programmers from VSO and SPASE for helping to characterize the issues as viewed by the scientists and data archives, and for putting up with my incessant questions.

## References

Auchère, F. (2002) *Preliminary Results from the SOHO Offpoint*. (Updated August 20, 200) [http://umbra.nascom.nasa.gov/eit/eit\\_guide/offpoint.htm](http://umbra.nascom.nasa.gov/eit/eit_guide/offpoint.htm)

Federal Geographic Data Committee (1998) *Content Standard for Digital Geospatial Metadata*. *FGDC-STD-001-1998*

Hill, F.; Martens, P.; Yoshimura, K.; Gurman, J.; Hourclé, J.; Dimitoglou, Ge.; Suárez-Solá, I.; Wampler, S.; Reardon, K.; Davey, A.; Bogart, R. S.; Tian, K. Q. (2009) The Virtual Solar Observatory—A Resource for International Heliophysics Research. *Earth, Moon, and Planets*, 104(1-4), 315-330. doi:10.1007/s11038-008-9274-7

Hourclé, J.A. (2007). FRBR in a Scientific Context. Science Archives in the 21st Century, Adelphi, MD - April 25-26, 2007.

wasn't physically published 'til 2009; need to adjust identifiers for references

Hourclé, J., Suárez-Solá, I., Davey, A., Tian, K., Yoshimura, K., Martens, P., Gurman, J., Hill F., and Bogart, R. (2007). Design Considerations for Data Catalogs, *Eos Trans. AGU*, 88(52), Fall Meet. Suppl., Abstract SH51A-0261

Hourclé, J.A. (2008a). "Reconciling Heterogeneous Data Catalogs". *21st CODATA International Conference*, Kyiv, Ukraine - October 7, 2008.

Hourclé, J.A (2008b), FRBR Applied to Scientific Data , *Proc. ASIS&T*, 45(1). doi:10.1002/meet.2008.14504503102

Hourclé, J.A. (2008c) Data Relationships: Towards a Conceptual Model of Scientific Data Catalogs. *Eos Trans. AGU*, 89(53), Fall Meet. Suppl., Abstract IN22A-03.

Hourclé, J.A. (2009a) Interoperability in the Space Sciences, 2009 ASIS&T, Vancouver, BC, November 11, 2009.

Hourclé, J.A (2009b) Ignored Issues in e-Science: Collaboration, Provenance and the Ethics of Data. *Eos Trans. AGU*, 90(52), Fall Meet. Suppl., Abstract IN31B-1008.

IFLA Study Group on the Functional Requirements for Bibliographic Records, & International Federation of Library Associations and Institutions. (1998). *Functional Requirements for Bibliographic Records: final report*. München: K.G. Saur.

King, J. (2007) personal communication.

Leibacher, J. (2009) Proposed Target Identification Convention for Solar Observations. *SolarNews*, 2009(15).

Metcalf, T. (2003) Correcting TRACE Pointing with MDI WL Images or EIT EUV Images. (Updated: September 26, 2003) <http://www.cora.nwra.com/~metcalf/TRACE/pointing.html>

Mink, D. (2002) *Proposed FITS Standard World Coordinate Systems*. (Updated: April 3, 2002) <http://tdc-www.harvard.edu/software/wcstools/wcstools.fits.html>

Norris, R., Andernach, H., Eichhorn, G., Genova, F., Griffin, E., Hanisch, R., Kembhavi, A., Kennicutt, R. and Richards, A. (2006) Astronomical Data Management, *Highlights of Astronomy, Volume 14*, ed. K.A. van der Hucht. arXiv:astro-ph/0611012

National Aeronautics and Space Administration (2005) *A Framework for Space and Solar Physics Virtual Observatories: Results from a Community Workshop sponsored by NASA's Living With a Star Program*. 27-29 October 2004, Greenbelt, MD.

National Aeronautics and Space Administration (2009) *NASA Heliophysics Science Data Management Policy, Version 1.1*. 12 April 2009.

National Aeronautics and Space Administration (2010) *Earth Science Data Terminology and Formats*. (Updated: January 11, 2010). <http://nasascience.nasa.gov/earth-science/earth-science-data-centers/earth-science-data-terminology-and-formats>

National Research Council (1982). *Data Management and Computation Volume 1: Issues and Recommendations*. Report of the Committee on Data Management and Computation, Space Sciences Board, National Research Council. National Academy Press.

National Research Council (1984). *Solar-Terrestrial Data Access, Distribution, and Archiving*. Report of the Joint Data Panel, Committee of Solar and Space Physics, Space Science Board, and of the Committee on Solar-Terrestrial Research, Board on Atmospheric Science and Climate, National Research Council. National Academy Press.

National Research Council (1997). *Massive Data Sets: Proceedings of a Workshop*. Committee on Applied and Theoretical Statistics, National Research Council. National Academy Press.

Sibeck, D.G. and Kucera, T. (2002) *Report and Recommendations of the LWS Science Data System Planning Team*, January 2002.

Suárez-Solá, F.I; Davey, A.R. and Hourclé, J.A. (2008) What Are We Tracking ... and Why? *Eos Trans. AGU*, 89(53), Fall Meet. Suppl., Abstract IN11C-1047.

Thompson, W.T. (2000) Standardized Coordinate Systems for Solar Image Data. *Proceedings of the 1st Solar and Space Weather Euroconference*, 25-29 September 2000, Santa Cruz de Tenerife, Tenerife, Spain. Edited by A. Wilson. Noordwijk, Netherlands: ESA Publications Division, 2000 xi, p.645-8.

Thompson, W.T. and Wei, K. (2010) Use of the FITS World Coordinate System by STEREO/SECCHI. *Solar Physics*, 261(1), 215-222. doi:10.1007/s11207-009-9476-9