

Building a Virtual Observatory for Heliophysics

R.D. Bentley, University College London

Introduction

Heliophysics explores the Sun-Solar System Connection and is a relatively new discipline. It generalizes of the study of "space weather" to the whole Solar System and spans several existing disciplines – solar physics, heliospheric physics, and planetary magnetospheric and ionospheric physics. The desire to solve science problems that span the disciplinary boundaries is now driving the need to provide integrated access to data from all the communities that constitute heliophysics.

To facilitate this we need to find ways to select related data through searches of metadata across the different domains. We also need to ensure that any results are presented in a form that does not require detailed understanding of each of disciplines involved. The virtual observatory paradigm is designed to meet these needs.

A difficulty is that the disciplines have evolved independently over decades and even centuries. There has been little or no coordination between them and each has different ways of describing, storing and exploiting data and different views on the need for standards. These problems must be addressed if a virtual observatory is to be established.

In addition, heliophysics sits in a boundary layer between two communities, astrophysics and planetary sciences: the astrophysics community is interested because a study of the sun and solar system can help in the understanding of stellar observations; the planetary science community (including Earth sciences) are interested because of the influence that the Sun can have on the environment on and around the planets. The two communities are very different to each other and a heliophysics virtual observatory must be aware of the needs of both communities and be sympathetic to their different approaches.

All in all this requires a re-evaluation of the capabilities provided within each domain, extensive discussions on many topics and some corrective action where necessary.

Metadata is Important

To facilitate a virtual observatory for heliophysics we need to examine the types of metadata that are required. There are many ways to group these, one is:

- Search metadata – *used to identify time intervals and sets of data of interest*
- Observational Metadata – *used to describe the observations, e.g. FITS header*
- Storage Metadata – *describes how the data are stored and accessed*
- Administrative metadata – *allows the system to exploit the available resources*

Search Metadata

In heliophysics we are interested in how an event on or near the solar surface can propagate through the solar system and affect planetary environments. We may also want to work backwards from effects near a planet and see if we can determine the cause.

Searches should identify interesting time intervals (and hence data sets) based on a combination of events, features, etc. as described by metadata and summary data in the different domains. The whole point of heliophysics is that we are interested in the effects of phenomena that are moving; the nature of the observed events and their timing are therefore strongly dependent on the location of the observer.

Much of solar physics is based on remote sensed observations. For this type of data, spatial information may be expressed in terms of the observing frame (i.e. pixels in an image) or with respect to the surface of a rotating body (i.e. the Sun). Until the launch of STEREO the location of the observer was almost ignored by the solar community (except for helio-seismology) since most observations were close to the Sun-Earth line.

Heliophysics is looking at the bigger picture and in-situ observations are also extremely important. In interplanetary space the position of the observer relative to the Sun is key to understanding in-situ observations; the time that event occurs is the time that a phenomenon affects (passes) the observer. When the in-situ observations are made on or near a planet, the position of the observer with respect to the planet (and its Sun-planet line) is also important.

Relating events that are defined from in-situ data to those on or near the Sun requires an understanding of how the phenomena propagate. Details of the velocity structure of CMEs and the solar wind are not easy to determine, however, using propagation models to combine the available information including event data from various locations could provide the information needed to address the science problems of heliophysics.

In principle this all sounds fairly straightforward, but when you actually try to combine event and feature metadata from the different domains you run into problems. Each community has created its own search metadata, but there are concerns about the quality and integrity of these. As a consequence, there are concerns about whether they are adequate to support the types of searches we would like to conduct.

If you examine the data and look at how the lists are created you realize that they are deficient in several areas. Consciously or unconsciously researchers take such deficiencies into account when working in their own domains. It is harder for machines to do this – if crucial information is missing the results of a search could be misleading.

An example is the list of Solar Proton Events maintained by NOAA. On 20 January 2005 there was an X1.7 flare that was extremely geo-effective and was also observed by with a ground-level neutron monitors – a GLE. Although it is described in the literature as the “most intense for 15 years” and the “most spectacular of the Space Age”, it does not actually appear in the list of proton events. This is because the flux of 10 MeV protons became elevated several days earlier (on 16 January) and had not dropped back below the threshold of $10 \text{ particles cm}^{-2} \text{ str}^{-1} \text{ s}^{-1}$. While the event list generated from the GOES soft X-ray monitors includes flares even if they are observed during the decay of a large long duration event, the list of protons events does not. Thus, a search of databases of events occurring near the Earth (and other planets) might show the effects of the 20 January event, but a correlation would not be made with something happening on the Sun unless the light-curves were also consulted.

The parameters listed in the soft X-ray flare list are not perfect either. The event times are truncated if a new flare occurs and it is difficult to determine the true duration of an event from the event list alone. In addition, the locations of many flares are not identified in the list and some significant brightenings in soft X-rays or the EUV never make it onto the lists since they barely register on whole disk monitors. These events could be relevant to the disruption of filaments that are often not flare related (cf: the flare myth); depending on their location on the Sun they could also be associated with geo-effective phenomena.

Flare lists produced by instrument teams often have gaps in them because of times when the instrument is off – at night, in the SAA or if there is a problem – but information on the cause is usually not easily accessible. As a consequence, a “blind” comparison with other types of event lists may result in spurious null results.

Although the examples given relate to solar physics, similar problems exist in all the disciplines involved in heliophysics. As it stands, the comparison of currently available event lists could give a distorted picture of what has occurred. Many of these problems could be fixed by reprocessing the source data and regenerating the event and feature lists bearing in mind that they may need to be compared to lists from other, very different sources.

Observational Metadata

Good observational metadata are essential when trying to compare data from one source with observations from other sources. For observations made from the Earth's surface, it is often necessary to combine observations from many time zones to obtain the complete coverage on an event; this means that you are subject to the vagaries of how several organizations handle metadata.

The space-based observatories are generally in better shape than their ground-based counterparts – since the instruments are often built by international consortia the provision of the data to a widely scattered community is planned from the outset and the data use usually better documented. This is of course not universally true – even a mission as well planned as SOHO has its problems with file headers.

Incomplete or wrong parameters associated with observational metadata are something that makes implementing virtual observatories difficult. Just because the data are more accessible, users expect things to be more perfect than is possible. While a virtual observatory can fix some of the problems itself, one of its principal goals should be to try to encourage the use of standards to improve the quality of the metadata in all quarters.

Storage Metadata

How data are stored in an archive can make a lot of difference to their accessibility. Existing virtual observatories have opinions on how easily data sets can be integrated depending on their organizational structure; passing this knowledge onto providers ought to generally improve access to data.

The European Grid of Solar Observations (EGSO) has the concept of resource-rich and resource-poor providers:

- Resource-rich providers are very capable and should be able to provide whatever is required in response to a simple query – an excellent example of this type of

provider is the Solar Data Analysis Center (SDAC) at NASA's Goddard Space Flight Center.

- Resource-poor providers may be able to do little more than to make the data accessible through the Internet – the virtual observatory may need to take additional steps to make the data usable.

Although one of the purposes of virtual observatories is to make data sets equally accessible from both resource-rich and resource poor providers, clearly encouraging best practices for data storage will improve the overall situation.

At the Virtual Observatories in Geoscience (VOiG) Conference in June 2007 we discussed whether we should produce general guidelines describing ways that data should be organized – not just for heliophysics but for all disciplines since this is a common problem. Doing this would be of assistance to resource-poor providers since they would have something that could easily be adopted; it should also be useful to archives that are being designed to support new observatories. This discussion is now being carried on within a relevant IAU working group¹.

Conclusions

There is now a strong desire to address science problems that span disciplinary boundaries. Also, the technology to achieve this is now available – recent advances in computing in the form of cheap storage and processing power coupled with the advent of the Internet mean that it is now extremely easy to share data. However, there are issues related to the data and metadata that need to be addressed.

Some of the tools developed using the EGSO Application Programming Interface (API) make it simple to determine the quantity, and to an extent the quality, of the data held by solar archives that have been integrated. The general conclusion is that observatories need to be strongly encouraged to make more data available and to improve the quality of the metadata that they provide to describe the data – the observational metadata.

The problems are not restricted to the solar domain – they exist in all domains. An underlying reason is that not all groups have been used to sharing their data and this has allowed deficiencies to creep in – this causes difficulties when searching across domains.

A virtual observatory for heliophysics needs to engage the communities in discussions on ways to improve the situation. This should be done primarily by improving the metadata – as far as possible it should not require any changes to existing data sets.

In May 2007 a proposal was submitted to the European Commission to establish a virtual observatory for heliophysics under FP7. A major component of the Heliophysics Integrated Observatory, HELIO, is intended to facilitate discussion with providers in order to address these problems and encourage the adoption of standards. If funded this will be a significant boost to the efforts started under the International Heliophysical Year.

¹ IAU Working Group on International Data Access, <http://www.mssl.ucl.ac.uk/grid/iau>