Future Generation Computer Systems 29 (2013) 2157–2168

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

HELIO: Discovery and analysis of data in heliophysics

Robert Bentley [a], John Brooke [b,*], André Csillaghy [c], Donal Fellows [b], Anja Le Blanc [b], Mauro Messerotti [d], David Pérez-Suárez [e], Gabriele Pierantoni [f], Marco Soldati [c]

[a] Mullard Space Science Laboratory, University College London, Holmbury St. Mary, Dorking, Surrey RH5 6NT, UK
[b] University of Manchester, Oxford Road, Manchester M13 9PL, UK
[c] Fachhochschule Nordwestschweiz, Institute of 4D Technologies, Steinackerstrasse 5, 5210 Windisch, Switzerland
[d] INAF-Astronomical Observatory of Trieste, Loc. Basovizza n. 302, 34012 Trieste, Italy
[e] Finnish Meteorological Institute, Erik Palménin Aukio 1, 00560 Helsinki, Finland
[f] Trinity College Dublin, College Green, Dublin 2, Ireland

HIGHLIGHTS

- Heliophysics has sparse datasets with rich data models and complex correlations.
- HELIO provides a database search and correlation services to support scientists.
- The Web user interface integrates access to databases and the Taverna Server.

ARTICLE INFO

Article history:
Received 5 March 2012
Received in revised form
18 February 2013
Accepted 6 April 2013
Available online 19 April 2013

Keywords:
Heliophysics
e-Science
Ontologies
Web services
Workflows

ABSTRACT

Heliophysics is the study of highly energetic events that originate on the Sun and propagate through the solar system. Such events can cause critical and possibly fatal disruption of the electromagnetic systems on spacecraft and on ground-based structures such as electric power grids, so there is a clear need to understand the events in their totality as they propagate through space and time. The e-Science challenge posed is that the data was gathered by many observatories and communities that have hitherto not needed to work together. Firstly, this involves the problem of helping users to more easily find and understand the relevance of data, especially data from outside their domain. Secondly, it involves solving challenges of data integration. We describe the design of the HELIO infrastructure, based on the use of Web services linked together by workflows and accessible via portal-based user interfaces. We also discuss current progress in the implementation of this infrastructure and the feedback from the user community.

1. An e-Science infrastructure for heliophysics

Heliophysics is the study of the effects of the Sun on the solar system; it addresses problems that span a number of existing disciplines—solar and heliospheric physics, and magnetospheric and ionospheric physics for the Earth and other planets. The discipline is closely related to the study of Space Weather (whose effects on modern technology are well documented [1,2]), but heliophysics is more generalised, covering all parts of the solar system rather than just the Sun–Earth connection.

In order to undertake searches that are scientifically interesting in heliophysics, we need to understand the origins of phenomena and how they propagate through interplanetary space, i.e., the path they follow and the time scales involved. This requires the ability to track both spacecraft and different types of solar emission in four dimensions, which is a key difference from other astrophysical searches based on images of the "deep sky" which can use a two-dimensional coordinate system based on the celestial sphere [3].

Virtual observatories have been a highly successful approach to issues of data sharing and reuse in astronomy [4]. A virtual observatory for heliophysics (VHO) needs extra tools to extend the essentially two-dimensional search space of deep-sky astronomy, since even though the deep-sky astronomy community has developed standards for data models and access methods that reduce the complexity of the e-infrastructure required for a virtual observatory, they do not address the more complex search problems of heliophysics. Within the Heliophysics Science Division of NASA,

* Corresponding author. Tel.: +44 1612750630.
E-mail addresses: b.bentley@ucl.ac.uk (R. Bentley), john.brooke@manchester.ac.uk (J. Brooke), andre.csillaghy@fhnw.ch (A. Csillaghy), donal.k.fellows@manchester.ac.uk (D. Fellows), anja.leblanc@manchester.ac.uk (A. Le Blanc), messerot@oats.inaf.it (M. Messerotti), dps.helio@gmail.com (D. Pérez-Suárez), gabriele.pierantoni@scss.tcd.ie (G. Pierantoni), marco.soldati@fhnw.ch (M. Soldati).

0167-739X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.future.2013.04.006

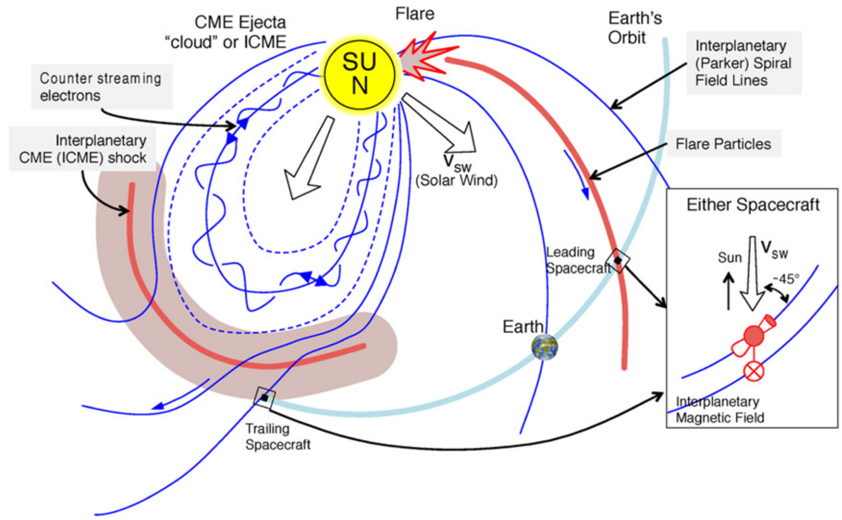

Fig. 1. Illustration of how the location of an instrument (using STEREO mission as exemplar) is a vital consideration for whether it is able to provide an observation relevant to the study of an event. The leading spacecraft will detect particles from a flare issuing from one side of the Sun, whereas the trailing spacecraft will detect a coronal mass ejection (CME) shock wave on the other side.

the approach has been to establish a number of small virtual observatories (often called the VxOs) that address individual parts of the heliophysics system—Sun, heliosphere, magnetosphere, etc. In the HELIO (Heliophysics Integrated Observatory)[1] project, we decided instead to address the complete system, since it allows a richer and more comprehensive search environment, and deal with the problems of data management separately.

The communities involved in heliophysics have evolved independently over decades, even centuries. Although the links between the effects observed in the disciplines are now evident, there have been virtually no attempts to coordinate the way the scientists collectively conduct their data analyses. As a consequence, there are considerable differences in the way the communities store, describe, and think about data, and this has a consequence of encouraging scientists in the domain to focus on extremely narrow datasets instead of looking at the much broader sweep of data available from the past 40 years of data collection; it is these challenges that the HELIO project was established to address.

In order to facilitate the study of this new discipline, HELIO needed to tackle issues in a number of areas related to two basic requirements.

- Provide integrated access to data from all the domains of heliophysics that are held in archives around the world.
- Provide the means to conduct searches across the domains to identify datasets of interest.

We have previously [5] described the scientific challenges involved. In the present work, we describe the e-Science infrastructure we are creating to meet these challenges. A major research problem is to search multiple catalogues or databases to track the development of an event when the effects of that event travel at different speeds. Heliophysical events are first observed (remotely) on the Sun, and then propagate through the solar system while potentially being detected by a variety of space-based and earth-based instruments. Effects caused by photon emissions require line-of-sight view of the source, and any delays are related to exactly predictable light travel times; those that are caused by particles occur with much longer delays. These delays are not exactly predictable, due to the interaction of the particles with the interplanetary magnetic field, and in most cases the effects are only experienced if the propagating phenomena directly passes the observer (see Fig. 1).

We have used previous work on virtual observatories as much as possible, and we have adopted a number of techniques and standards developed within the International Virtual Observatory Alliance (IVOA) [6]. However, the dynamic nature of heliophysics (in particular, its strong dependency on time series) has meant that we have had to borrow Web services approaches [7,8] from other fields where data sources change more rapidly than in deep-sky astronomy (e.g., biosciences). An example of this is our use of workflows to link data sources and our workflow repository that allows queries to be re-run as the data in the data services changes.

In Section 2, we describe the e-Science challenge of heliophysics in more depth through a case study. In Section 3, we describe how we meet the challenges of cross-catalogue searches with non-trivial relationships between search regions and multiple data models. In Section 4, we describe the architecture we have built. In Section 5, we describe the scientific interface to the HELIO virtual observatory. In Section 6, we summarise the wider impact on e-Science and how our methods will respond to technology developments (e.g., Cloud Computing).

2. Scientific case studies

2.1. Scientific setting

One of the key subjects studied by heliospheric physics is the release of large amounts of ionised particles, called plasma, that propagate through the heliosphere and interact with planetary environments. Particles are accelerated by large solar explosions called flares or by prominences that erupt and cause the ejection of "blobs" of plasma into interplanetary space. They are commonly known as coronal mass ejections (CMEs) [9]. Another phenomenon under study is called the stream interaction region (SIR); such a region forms when a fast solar wind stream overtakes a slower one [10]. SIRs are representatives of a variety of perturbations in the solar wind ambient plasma. Both phenomena, CMEs and SIRs, originate by events produced at the Sun.

The analysis of these and similar phenomena involves the availability of multi-instrument and multi-wavelength data sampled at multiple locations. It also requires suitable propagation models so that it is possible to track the temporal and spatial evolution of

[1] http://helio-vo.eu.