HELIO: Discovery and Analysis of Data in Heliophysics

Robert Bentley*, John Brooke†, André Csillaghy‡, Donal Fellows†, Anja Le Blanc†,
Mauro Messerotti§, David Pérez-Suárez¶, Gabriele Pierantoni¶, Marco Soldati‡
* Mullard Space Science Laboratory, University College London, Holmbury St. Mary, Dorking, Surrey RH5 6NT, U.K.
† University of Manchester, Oxford Road, Manchester M13 9PL, U.K.
‡ Fachhochschule Nordwestschweiz, Institute of 4D Technologies, Steinackerstrasse 5, 5210 Windisch, Switzerland
§ INAF-Astronomical Observatory of Trieste, Loc. Basovizza n. 302, 34012 Trieste, Italy
¶ Trinity College Dublin, College Green, Dublin 2, Ireland
Email: john.brooke@manchester.ac.uk

***Abstract*—Heliophysics is the study of highly energetic events that originate on the sun and propogate 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 propogate through space and time. This poses a fascinating eScience challenge since the data is gathered by many observatories and communities that have hitherto not needed to work together. We describe how we are developing an eScience infrastructure to make the discovery and analysis of such complex events possible for the communities of heliophysics. The new systematic and data-centric science which will develop from this will be a child of both the space and information ages.**

I. 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 charged particles in 4-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 (VxO for short) have been a highly successful approach to issues of data sharing and re-use 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 VxO, they do not address the more complex search problems of heliophysics.

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 space physicists store, describe and think about data, and this has a consequence of encouraging scientists in the domain to focus on extremely narrow data-sets 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 set up to address.

In order to facilitate the study of this new discipline, HELIO needs 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 data-sets of interest.

We have previously [5] described the scientific challenges involved. In the present work we describe the eScience 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- 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 use previous work on VxOs as much as possible. However, the dynamic nature of heliophysics (in particular,

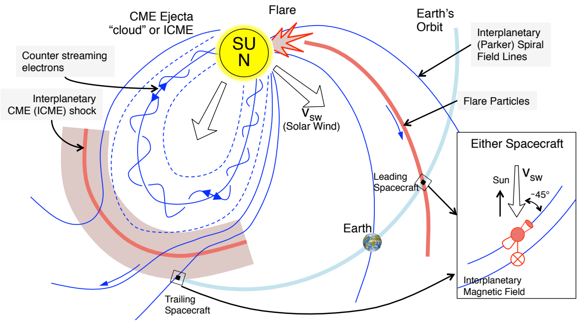

Fig. 1. Illustration of how the location of an instrument (using STEREO mission as exemplar) is a vital consideration for whether it is able to provide an observation relevant to the study of an event. The leading spacecraft will detect particles from a flare issuing from one side of the sun, whereas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

its strong dependency on time series) has meant that we have had to borrow Web Services approaches from other fields (e.g., biosciences). Thus we present a case-study of a fascinating cross-over between e-Science techniques evolved from different disciplines but resulting in a common approach to infrastructure building. In Section II, we describe the e-Science challenge of heliophysics in more depth through a case study. In Section III, we describe how we meet the challenges of multiple data models and so enable cross-catalogue searches. In Section IV, we describe the architecture we have built. In Section V, we describe the scientific interface to the HELIO VHO. In Section VI, we summarise the wider impact on eScience and how our methods will respond to technology developments (e.g., Cloud Computing)

II. Scientific Case Study

A. *Scientific setting*

One of the key things studied by heliospheric physics is the release of large amounts of ionized 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. Another phenomenon under study is called stream interaction regions (SIR) [6] where different regions of the solar wind travel at different speeds. This causes shock fronts in the solar wind. SIR are representatives of a variety of perturbations in the solar wind ambient plasma. Both phenomena originate at the Sun.

The analysis of these and similar phenomena involves the availability of multi-instrument, multi-wavelength, and multi-point data. It also requires suitable propagation models so that it is possible to track the temporal and spatial evolution of phenomena with respect to triggering events (e.g., solar flares), interaction events during interplanetary propagation (e.g., particle beam acceleration and reflection at the shock front), and interaction processes with planetary magnetospheres and atmospheres (e.g., compression and energy transfer, particle injection). Hence, a scientist who wants to perform such an analysis has to: 1) identify the ancillary heliospheric events that have concurred to determine the observational scenario focused on the primary heliospheric event of interest; 2) identify the data sources; 3) run propagation models to generate a time frame relevant to the occurrence of the various events; 4) download the data sets; 5) carry out the data integration; 6) perform the physical modeling and interpretation.

Ground-based and space-based data are usually needed, where *in situ*, multi-point and multi-spacecraft observations play a fundamental role in characterising the propagating heliospheric event. Such data are typically stored in dedicated archives, that can be accessed via web as standalone facilities or through portals which provide a common interface to different archives. There are VxOs that provide access to a variety of data sets, but they are usually sub-domain specific, i.e., they focus only on data relating to solar physics, space physics, magnetospheric physics, etc. Hence, the scientific user has had to manually follow with the workflow outlined above, which makes the preparatory phase of his/her research in heliophysics quite demanding.

B. *Case study: tracing a CME by auroral storms*

As a sample case study aimed at showing the improvement in research workflow that HELIO will be able to provide, in the following we briefly examine the study of an interplanetary shock traced by planetary auroral storms from the Sun to Saturn [7].

Auroral storms are associated with the perturbation of planetary magnetospheres stressed by the arrival of a coronal mass ejection (CME) at the Earth (see Fig. 2) and an ICME (interplanetary CME) at farther planets, after having travelled to 1 AU and beyond, respectively, with time-of-flight increasing with distance, and ranging from tens to hundreds of hours according to the propagation speed as well.

The data from SOHO (SOlar and Heliospheric Observatory, a NASA and ESA spacecraft located at the L1 Lagrangian point), in particular the instruments LASCO, EIT and VIRGO, allowed to characterise the CME observed at a certain date and time. The data from POLAR (a NASA spacecraft for probing the Earth magnetosphere) provided information on perturbation and auroral activity observed at 1 AU. In turn, the integration of these data provided information on the CME dynamics. The CME parameters near the Earth have been derived by the data from the Solar Wind Experiment of the WIND spacecraft, and from the Advanced Composition Explorer.

A magnetohydrodynamic model of the CME was fed with the observed CME parameters, and this allowed to estimate the arrival time of the ICME at Jupiter and Saturn, respectively. Observational data from the RPWS instrument aboard Cassini and data from Galileo allowed to identify the CME-associated plasma shock at Jupiter, whereas auroral activity could be identified first on Jupiter and later on Saturn from images taken by the Hubble Space Telescope.

As this example shows, data from at least nine repositories had to be handled, i.e., identified placing searches based on

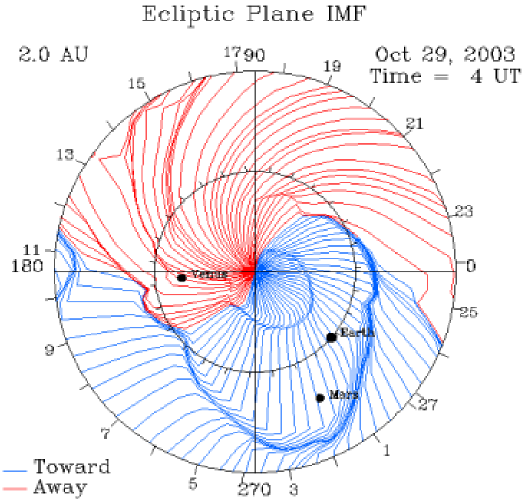

Fig. 2. Instantaneous snapshot of propagation model for charged particles, showing distortions to the interplanetary magnetic field in the plane of the solar system caused by coronal mass ejections (with thanks to J. Luhmann). Of particular note are the large distortions (from the close-to-ideal pattern in the top half of the figure) in field line directions due to the CME.

dates and times based according to a propagation model, and properly integrated in a common physical framework to carry out a global modeling of the observational scenario. This process is rendered tricky due to the intrinsic properties of the data, and the lack of full standardization of the data descriptors due to them evolving in heliophysics since the beginning of explorations of interplanetary space.

It is within this framework that HELIO provides the scientist with an operational scenario for heliophysical data handling. It relieves him or her from the burden of data source identification and data integration, as its web interface makes it possible to place complex searches on multiple data repositories relevant to heliospheric data in a unified, user-transparent way, as reported in Section V. This greatly facilitates the research, and creates a favourable operational environment for knowledge discovery.

III. Semantics in Data Models and Ontologies

Heliophysicists use a variety of data formats, data dictionaries, and data models in their work. Depending on their special area of interest it is difficult to understand and use data products created by a group outside of that area. A lot of these data products do not contain enough metadata to enable scientists to interpret them without the assumed knowledge within the source community. This is important because the HELIO infrastructure is not the only way for scientists to find and work with heliophysical data; different organisations (e.g., NASA, ESA) set up VxOs in which researchers can perform a subset of the functionality provided by HELIO. Which subset they cover depends on the concrete speciality these virtual observatories were designed to perform.

HELIO is designed to cross the boundaries between subcommunities within heliophysics. The project strives to provide services that can be used on their own but are also easy to integrate with one another. The data provided is encoded in the VOTable [8] format, an XML representation of tabular data that is widely used in the astronomy and astrophysics communities, which allows it to be rich in metadata, contain a full provenance trail, and be annotated using the community standards UCD [9] and UType. The UType attribute in particular provides a reference into a data model, which is key for driving semantic matching.

We are also aiming to bridge the gap arising by different data standards in different communities within heliophysics by creating an ontology that maps terms from these data standards to each other.

A. *Data model*

The services in the HELIO system (see Section IV) are designed to work on their own, but in order to perform more complex tasks the user needs to execute these services in combination with each other, using the outputs of one service in defining the criteria for a call to another. We have created an overarching data model in which we define the content of the services, ensuring that the data resulting from one query will be semantically compatible with subsequent queries.

All HELIO services produce their output in VOTable format, which was designed for the exchange of data in tabular form by the IVOA [10] (the standard body for virtual observatories). It is not specific to any content and does not contain any requirements of the description of the content. Even though the format can be read by a number of applications, only the author of such a file can make sure the content is meaningful to the recipient; this is enabled through the presence of two hooks in the VOTable format that are used to attach semantic content to the data. The first one is the “UCD”, which is a list of terms from a controlled vocabulary, though the current version of UCD, 1.23, does not contain appropriate terminology for heliophysics and the terms that are included do not provide the granularity required for a satisfactory mapping between tables. Since VOTable version 1.1, we can provide references into our own data models through the use of the second hook, the “UType”. That means we can create a data model describing the data exactly to the level of detail required to enable the use of the content between services.

An analysis of the existing data models in heliophysics showed that there are few well-defined data models, of which the most widely used data model is SPASE [11]. However, SPASE defines the structure of the data but does not deal with the meaning of the content, and the resulting UType references in a VOTable would be meaningless.

The HELIO data model was therefore constructed *de novo* with the idea to represent the semantics of the underlying data. Fields with the same content should have the same UType tag in the VOTable no matter which service has produced them. The resulting data model should not only be usable by the HELIO services, but should be well enough structured to be also easily usable by other community data providers. By creating a new data model, we run the risk that it will not be

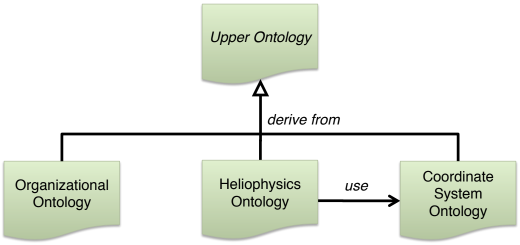

Fig. 3. The structure of the ontologies, showing how they relate to each other with an upper ontology providing common concepts allowing the ontologies to be used together.

used outside of this project, but our hope is that by tackling the semantics of the data in ways that other data models did not do, we can reach adopters outside our group.

B. Ontology

Different parts of the heliophysics community use different data models, data dictionaries, and file types to describe, store and exchange their data. These different data products were developed completely independently of each other, and often use different keywords to describe what the same thing is physically.

We addressed this through creating an ontology describing the whole heliophysics discipline. In the first stage of this process, we created an "Upper Ontology" that contains the basic concepts used in heliophysics, so creating a semantic skeleton for the science. It is logically structured (see Fig. 3) in a way which makes maintaining the parts easier and consists of:

Organizational Ontology This contains the structure and properties of data, infrastructure and people.

Coordinate System Ontology This contains classifications of coordinate systems, and parameters relating to coordinate systems.

Heliophysics Ontology This contains the domain concepts of this community. It uses the Coordinate Systems Ontology.

Upper Ontology This includes all concepts of the previous three ontologies, and adds properties which bind concepts of the different sub-ontologies together.

In the second stage, we used the "Upper Ontology" to map terms from the SPASE data model, the PDS data dictionary [12], the EGSO data model [13], and the HELIO data model onto that structure. We created individual annotation types for each of the data products and, where appropriate, hierarchies of annotation types. This enables us to use different levels of detail in the integration. Annotations have the advantage that they can be created for both classes and individuals, but a problem is that common ontological tools can't reason over them, which needs to be considered when queries are constructed. The resulting ontology (see Fig. 4) can be queried for terms in these different data products that represent the same concept or a related higher or lower level concept. Of course, the ontology can only provide these terms for areas where these data products actually cover the same ground, such as in the terms used for spacecraft names. Beyond the mapping of terms the ontology can also inform about the concepts used in heliophysics and their relation to one another.

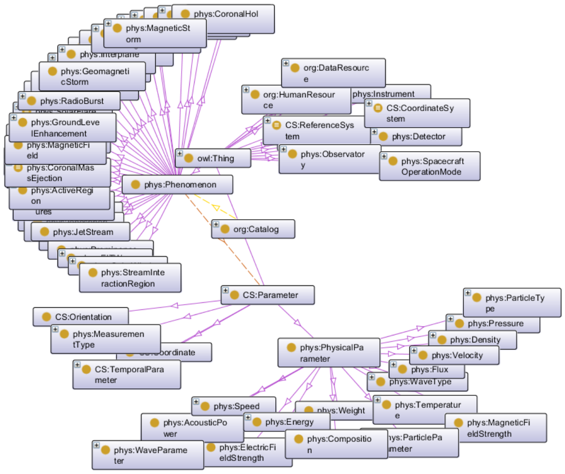

Fig. 4. The classes in the HELIO ontology, together with selected types of inter-class relation.

The ontology is integrated into a Semantic Mapping Service, which is a web service providing a SOAP interface to its functions (a part of the metadata category described in Section IV-B, though not a user of the HQI due to the results being non-tabular). This service allows the ontology to be integrated into other the parts of the HELIO system, or to provide semantic mappings to workflows.

IV. The Architecture of the HELIO infrastructure

A. Overall architecture

The HELIO infrastructure is based on the concepts of a Service Oriented Architecture [14]. SOAs feature a set of loosely coupled components, and so have two main advantages for HELIO:

1) The components can be deployed redundantly at different locations, increasing the overall stability of the system.
2) The components can be developed independently at different locations by different teams, so supporting the distributed nature of the project consortium.

Another key aspect of the use of a SOA is that the primary data resources (notably images and spectra) are large and generated at a high rate[1] so keeping the catalogues describing that data close to the depository institutions minimizes the number of large transfers that need to occur. This naturally leads to distributed catalogues due to the substantial number of organizations participating in heliospheric-related research.

The other principles used in the design of the HELIO architecture [16] were that services should, to as great an extent as possible, permit multiple access methods (minimally

[1] A *single* space-based observatory such as the SOHO satellite can produce 0.5Gb/day continuously [15]. Ground-based observatories can have much higher data rates. There are over 50 observatories, with over 200 instruments producing many different types of data, and with collection happening over many decades; the oldest complete datasets start in the nineteenth century.

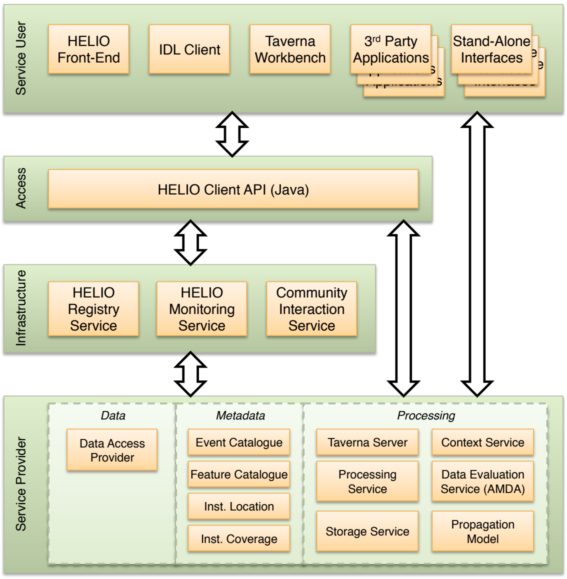

Fig. 5. Structural view of the main components involved in the HELIO infrastructure. The arrows denote communication flows.

including both integrated and standalone modes), that no particular client or workflow system should be specially favoured, that the system should respect the security policies of the service providers to as great an extent as possible (a real issue when some processing services require significant resources to operate), and that as much non-scientific information as possible should be hidden from the scientific users (i.e., that technical details of logins, data location, etc. should be kept shrouded unless specifically requested). A final key principle was that the technology used should *not* be at the bleeding edge; the focus of the project is entirely on providing a production-ready technology platform to support the science.

Fig. 5 shows the conceptual architecture of the components involved in the HELIO infrastructure. The diagram is divided into four main areas.

Service Provider Components that implement services providing access to data, metadata, on-demand processing and storage capabilities.

Infrastructure Components that are required for management, maintenance and security handling of the HELIO infrastructure. Consumed principally by the access layer.

Access Responsible for connecting and integrating the underlying services and for facilitating access to the infrastructure for different "service user" components. This layer handles security, failover, service resolution, etc.

Service User Components that provide the interface between human beings and the underlying infrastructure.

Although in simple usage, information proceeds along the major flow directions identified in Fig. 5, this is not the only way in which things can work. For example, Fig. 6 shows more complex interactions that can exist when a workflow server is in use. As can be seen, users can connect to a centralized Graphical User Interface (the HFE, see Section V) that uses an instance of the Taverna Server, as well as using a local instance of a Taverna Workbench to define workflows. They can also access the HELIO API directly through Java [17] or IDL (Interactive Data Language [18]) code. Finally, some of the services also offer standalone graphical user interfaces that offer advanced functionalities not available in the HFE.

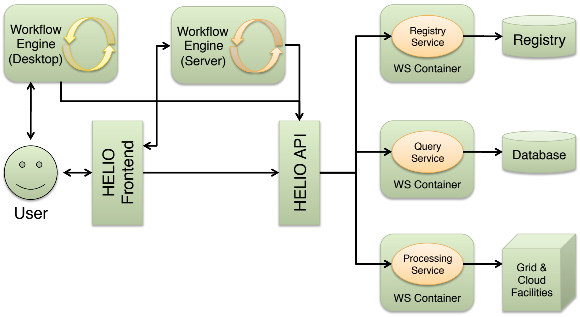

Fig. 6. Information flows in complex HELIO usage, showing how a workflow engine may be used within the front-end, or instead of it.

B. Service provider components

The service provider components are divided into three categories: data, metadata and processing.

The *data* category contains the Data Provider Access Service (DPAS), which provides uniform access to a multitude of archives with data about heliospheric observations. The DPAS implements connectors to various types of archives, such as FTP- and HTTP-archives, web services, relational databases and virtual observatories outside the heliophysics field.

The *metadata* category contains services for accessing secondary catalogues and other types of metadata. These catalogues include collections of events observed on the Sun and in the heliosphere, features of the Sun as they evolve over time (e.g., filaments, active regions, coronal holes), and descriptions of what instruments were observing and where they were located at the time.

Access to the metadata catalogues is given primarily through the HELIO Query Interface (HQI), a common standard interface for catalogue queries that supports both REST and SOAP styles of use; service users may use either to get the same results. Conceptually, it supports a parametric query style to query tabular data; parametric queries are best suited to express and implement cases where the data model is sufficiently well defined. The results of the queries are formatted as VOTables.

The *processing* category holds services for on-demand processing of data or metadata. Depending on the scientific question asked to HELIO some information cannot be prepared in advance but has to be computed based on given parameters. HELIO provides several types of processing components:

Taverna Server A workflow engine suited to combine multiple services into a more complex workflow (see section IV-F).

HELIO Processing Service The HPS provides access to a high performance computing infrastructure, used for resource intensive data processing such as image analysis.

HELIO Storage Service The HSS acts as utility service for the HPS and other services to store large result sets.

Context services These give access to predefined plotting services, e.g., to create a timeline plot of solar activity for a given date range.

Data Evaluation Service This acts as interface to the Automated Multi Dataset Analysis (AMDA) infrastructure [19], which provides a collection of tools to access and analyse heliophysical data.

Propagation Model This simulates the propagation of the effects of a solar event through space and time, allowing the scientist to relate observations made at different locations and time to the same event.

C. Infrastructure components

The infrastructure category provides helper services for the management of the HELIO infrastructure; these components are usually transparent to the end user and do not provide any information of scientific value.

The *HELIO Registry Service* (HRS) is a directory service to enable service discovery. Additionally, it provides information on how to use the services. The *HELIO Monitoring Service* (HMS) monitors the system by frequently polling the status of the service. In combination with the HRS it provides the failover and load balancing capabilities of the infrastructure.

The *Community Interaction Service* (CIS) implements the basement for authentication in HELIO. Moreover, it manages user profiles in a central place.

D. Access component

Depending on their needs client applications may choose to directly access individual HELIO services or they may use the Java-based HELIO API. The HELIO Java API facilitates access to the system by shielding users from the underlying infrastructure. It offers: 1) transparent discovery of services, 2) load balancing and failover (through the use of the HRS and HMS), 3) automatic handling of security and user profile management, 4) client stubs to access different service providers in a uniform way, and 5) utilities to combine services to solve more complex tasks.

E. Service user components

We support user access to HELIO services through multiple methods. The principal ones are:

HELIO Front-End The HFE provides an integrated browser-based interface to the HELIO services; it allows users to perform common searches and data retrieval actions in a user friendly way (discussed in more depth in Section V).

Taverna Workbench This allows users to define custom workflows for their own specific scientific use cases using a visual composition and configuration environment. It also supports the sharing of these workflows through social media [20].

HELIO IDL Client This enables access to HELIO through the IDL scripting environment. With IDL, users can interactively communicate with the HELIO system, and in this way combine the HELIO capabilities with advanced data analysis tasks.

Stand-alone Interfaces Most HELIO services have their own stand-alone interfaces, allowing them to be directly used over the web without integration with any other system.

We also support third-party application access, so that other virtual observatories, data warehouses, and graphical client applications can integrate with HELIO.

F. Workflows in HELIO

HELIO uses Taverna [21] [22] as its exemplar workflow system, as it provides a relatively simple mechanism for orchestrating multiple services into a single unit of processing. In addition to supporting the use of the Taverna Workbench, HELIO has an installation of Taverna Server which can execute workflows that have been created through the Workbench and stored in the myExperiment workflow repository [20] [23]. These stored workflows[2] are annotated with metadata that enables them to be automatically exposed to users through the HFE, enabling workflow use and reuse without the users having to install a complex piece of software like the Workbench. HELIO also enhances the Taverna Server installation with knowledge of HELIO's registry service and security model, so that workflows may access resources while only knowing the functionality they seek to use (e.g., access to a particular catalogue) and may use computation and storage services with the credentials of the user who invoked the workflow.

The workflows are principally comprised of processing elements that access HELIO's services (especially the query interfaces) via SOAP method invocations, interleaved with extra processing elements to extract and combine results. One example of this (see Fig. 7) is a composite query which takes a time period identified by the invoking user, during which they want to search for correlated features (coronal holes, etc.) and events (e.g., X-Ray flares) originating from the same part of the sun (i.e., within a certain distance across the sun's disk).

Architecturally, Taverna Server is a web service that is hosted within a Java web container. The server provides job and file management where the jobs are specialized to executing workflows created by the Taverna Workbench. Workflows are executed in a different local user account for each distinct user through the use of an impersonation module, allowing for the application of per-user security policies and accounting. (Because of the necessary use of impersonation to achieve this execution model, the server installation is not shared with any other services.)

V. The user interface

A. The design challenge

The major challenge of the user interface has been to support the concurrent selection and combination of data from many instruments so as to support heliophysics research such as described in Section II. Though many user interfaces exist already to perform single steps of the study, it is generally

[2] There are 34 shared workflows at the time of writing.

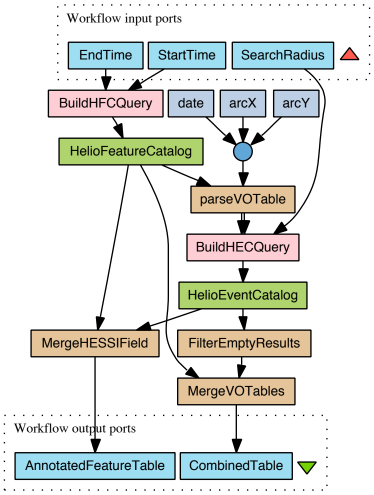

Fig. 7. Taverna workflow for querying the HELIO Feature Catalogue and HELIO Event Catalogue and merging the results. [24]

recognized that none of them provide an easy way of composing different analyses to answer larger research questions. Former projects, such EGSO [25], did provide an embryonic integrated user interface, but the previously existing systems did not fully support the way scientists want to work. As such, a major rethink was in order to increase the acceptance of the HELIO user interface approach.

This HELIO approach is to present the underlying capabilities as interface modules within a web portal with a common look and feel, where the modules can be chained together as the user requires based on the common underlying data model, with the underlying Web Services being accessed transparently by SOAP. The user is not made aware of the nature of the interaction with the services; the principal focus of the user interface is on the catalogue entries, images and spectra that constitute the scientifically-relevant parts of heliophysics. The approach also provides for customisation of the presentation to the focus of the communities that make up heliophysics, allowing the modules to be adapted to use updated algorithms and additional data sources, or even the restriction of the presented interface to just a subset of modules.

The user interface was made available to our testers early in the development process, and frequently updated according to user feedback. We have also promoted the adoption of the interface by the wider heliophysics community, since their experience and feedback are essential for ensuring that we support the methods of working of such groups. This is essential to ensuring that HELIO produces a sustainable suite of practical working tools for new scientific discovery.

Two of the major challenges for the design are:

- The addressed user groups and their needs (and expectations) are highly heterogeneous. Therefore, the user interface has to provide multiple different types of search task (e.g., searching by time and searching by location), and both user-guided and system-guided interaction styles.
- There are many heterogeneous data sets from different repositories, as well as complex search results, and their relationships have to be represented in a comprehensible way. Following Shneiderman’s ”Visual Information seeking Mantra” [26], we use representation techniques supporting overview, zoom/filter and details-on-demand visualisation that differ from the conventional user interfaces previously used in the heliospheric domain.

Therefore, the user interface provides techniques for the search tasks (where the focus is on how to present an input search range, how to manage the result, and how to pick a subset of the results) as well as tools to navigate through the data sets (with a particular focus on interactive visualization of the data sets through fetching appropriate previews).

B. Implementation

The user interface is realised as the *HELIO Front End* (HFE), a Rich Internet Application [27] written in Javascript [28] that provides access to the underlying HELIO API (see Section IV-D) and which mimics much of the capabilities that would be expected in a desktop application through the use of AJAX techniques [29]. Although this implies a certain development overhead compared to the use of technology such as Flash [30], it relieves users from installing browser plugins and increases the availability of the application to ordinary users. It also tries to minimize the use of novel user interface design elements by following general best practice as much as practical.

The HFE is centered around the data — which may either originate from catalogs within the system or from an uploaded VOTable — and the tasks performed on it. This distinguishes the HFE from both normal web applications, which are more workflow-oriented, and traditional scientific systems, which are function-oriented. In a function-oriented approach, input data is feed into a function, processed and new data is generated; this is comparable to a traditional scientific data analysis system, where the users need to know the details of the data, apply a function to it, and exactly know what they can expect back, but where there is no knowledge *in the system* of the nature of the input data or results.

By contrast, in a task oriented system, the data processing is done at an abstract level from the user’s perspective. This means that for a given data product, the user is presented with a set of tasks that can be applied to this data. These tasks are presented in natural language like: “Get observations for a given time range”, “See what instruments covered this period”, etc. This task-oriented approach supports the novice users to perform common tasks without having deep knowledge of the detailed science, allowing them to perform many analyses without having to learn the system in depth, while not preventing more advanced users from working with the data. This is supported through the use of simple data management tasks which retrieve the data, join data tables, and store the data products.

The input data to a task may consist of: 1) manually specified data, 2) data coming from a HELIO service such as the event catalogue or feature catalogue, or 3) data from an external source, such as a VOTable created by some scripting language. Most tasks are mapped to a query or processing service that runs in the HELIO infrastructure, and more advanced tasks may access workflows in the "helio" group in myExperiment repository (so allowing the scientific community to provide more functionality without developer intervention) and which are executed on a Taverna Server instance. The output data product generated by the task is either a VOTable document or a FITS image [31], and may be used as input to further tasks or downloaded to the users' local system for longer-term storage or specialist analysis.

VI. Conclusions and future work

The HELIO infrastructure is largely complete and key use cases are being deployed. The community consultation is proceeding via a series of workshops in which the requirements of the heliophysicists are being mapped onto the services and the services are linked in dynamic workflows that execute across a back-end infrastructure that transparently uses Grid and Cloud resources. The workflows represent a key resource for the community, just as they do in other disciplines and are shared via the myExperiment repository [20]. We have been careful to reuse previous work by the Virtual Observatory and eScience communities, and we believe that our success in doing this is a mark of the progress of eScience to becoming a more mature field of research. As a result, we are able for the first time to address the whole nature of heliophysics.

Studying the heliophysics discipline in a systematic manner will bring new challenges, and methods will be developed that can be applied in other data-centric sciences. Future work will integrate the data gathered from observations with models of the energetic processes of interplanetary space, allowing for example the models to be continuously calibrated with data in a similar manner to data ingestion in weather forecasting.

Acknowledgment

We would like to thank the Geophysical Institute, University of Alaska, Fairbanks, Alaska for the permission to use Fig. 2. This work was funded by the European Commission as part of the Seventh Framework Programme, Project No. 238969.

References

[1] J. Kappenman, L. Zanetti, and W. Radasky, "Geomagnetic storms can threaten electric power grid," *Earth in Space*, vol. 9, no. 7, pp. 9–11, 1997.

[2] R. Langley, "GPS, the Ionosphere, and the Solar Maximum," *GPS World*, vol. 11, no. 7, pp. 44–49, 2000.

[3] W. Fricke, "Definition of the celestial reference coordinate system in fundamental catalogues," in *IAU Colloq. 26: On Reference Coordinate Systems for Earth Dynamics*, vol. 1, 1975, pp. 201–222.

[4] A. Szalay and J. Gray, "The world-wide telescope," *Science*, vol. 293, no. 5537, pp. 2037–2038, 2001.

[5] R. Bentley, A. Csillaghy, J. Aboudarham, C. Jacquey, M. Hapgood, K. Bocchialini, M. Messerotti, J. Brooke, P. Gallagher, P. Fox *et al.*, "HELIO: The Heliophysics Integrated Observatory," *Advances in Space Research*, 2010.

[6] L. Jian, C. Russell, J. Luhmann, and R. Skoug, "Properties of stream interactions at one AU during 1995–2004," *Solar Physics*, vol. 239, no. 1, pp. 337–392, 2006.

[7] R. Prangé, L. Pallier, K. Hansen, R. Howard, A. Vourlidas, R. Courtin, and C. Parkinson, "An interplanetary shock traced by planetary auroral storms from the Sun to Saturn," *Nature*, vol. 432, no. 7013, pp. 78–81, 2004.

[8] F. Ochsenbein, R. Williams, C. Davenhall, D. Durand, P. Fernique, R. Hanisch, D. Giaretta, T. McGlynn, A. Szalay, and A. Wicenec, "VOTable: Tabular data for the Virtual Observatory," in *Toward an International Virtual Observatory*, ser. ESO Astrophysics Symposia, P. Quinn and K. Górski, Eds. Springer Berlin / Heidelberg, 2004, vol. 30, pp. 118–123, 10.1007/10857598_18. [Online]. Available: http://dx.doi.org/10.1007/10857598_18

[9] A. Martınez, S. Derriere, N. Gray, R. Mann, J. McDowell, T. Mc Glynn, F. Ochsenbein, P. Osuna, G. Rixon, and R. Williams, "The UCD1+ controlled vocabulary," *IVOA Semantics WG Recommendation*, 2005.

[10] M. Ohishi, "International Virtual Observatory Alliance," *Proceedings of the International Astronomical Union*, vol. 2, no. 14, pp. 528–529, 2006.

[11] J. Dickinson *et al.*, "SPASE collaboration," *Nucl. Inst. Meth. A*, vol. 440, p. 95, 2000.

[12] S. K. McMahon, "Overview of the Planetary Data System," *Planetary and Space Science*, vol. 44, no. 1, pp. 3–12, 1996, planetary data system. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0032063395001018

[13] K. Reardon, R. Bentley, M. Messerotti, and S. Giordano, "A solar data model for use in Virtual Observatories," in *Bulletin of the American Astronomical Society*, vol. 36, 2004, p. 796.

[14] R. Perrey and M. Lycett, "Service-oriented architecture," in *Applications and the Internet Workshops, 2003. Proceedings. 2003 Symposium on*. IEEE Computer Society, 2003, pp. 116–119.

[15] V. Domingo, B. Fleck, and A. Poland, "The SOHO mission: an overview," *Solar Physics*, vol. 162, no. 1, pp. 1–37, 1995.

[16] G. Pierantoni, B. Coghlan, and E. Kenny, "The Architecture of HELIO," in *Cracow Grid Workshop*, 2010.

[17] J. Gosling, *The Java language specification*. Prentice Hall, 2000.

[18] B. Stern, "Interactive Data Language," in *Proceedings of Space 2000: the Seventh International Conference; Albuquerque, NM*. American Society of Civil Engineers, 1801 Alexander Bell Drive, Reston, VA, 20191-4400, USA,, 2000.

[19] C. Jacquey, V. Génot, E. Budnik, R. Hitier, M. Bouchemit, M. Gangloff, A. Fedorov, B. Cecconi, N. André, B. Lavraud *et al.*, "Amda, automated multi-dataset analysis: A web-based service provided by the CDPP," *The Cluster Active Archive*, pp. 239–247, 2010.

[20] C. Goble and D. De Roure, "myExperiment: social networking for workflow-using e-scientists," in *Proceedings of the 2nd workshop on Workflows in support of large-scale science*. ACM, 2007, pp. 1–2.

[21] D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. Pocock, P. Li, and T. Oinn, "Taverna: a tool for building and running workflows of services," *Nucleic acids research*, vol. 34, no. suppl 2, p. W729, 2006.

[22] W. Tan, R. Madduri, A. Nenadic, S. Soiland-Reyes, D. Sulakhe, I. Foster, and C. Goble, "Cagrid workflow toolkit: A taverna based workflow tool for cancer grid," *BMC bioinformatics*, vol. 11, no. 1, p. 542, 2010.

[23] W. Tan, J. Zhang, and I. Foster, "Network analysis of scientific workflows: A gateway to reuse," *Computer*, vol. 43, no. 9, pp. 54–61, 2010.

[24] A. Le Blanc and D. Fellows, "Associate hessi flares with active regions," In myExperiment repository, http://www.myexperiment.org/workflows/2181.html, June 2011.

[25] R. Bentley, A. Csillaghy, and I. Scholl, "The European grid of solar observations," in *Proceedings of SPIE*, vol. 5493. Citeseer, 2004, pp. 170–177.

[26] B. Shneiderman, "The Eyes Have It: a task by data type taxonomy for information visualizations," *Visual Languages, IEEE Symposium on*, vol. 0, p. 336, 1996.

[27] M. Driver, R. Valdes, and G. Phifer, "Rich Internet Applications are the next evolution of the Web," *Gartner Research*, 2005.

[28] D. Flanagan, *JavaScript: the definitive guide*. O'Reilly, 1998.

[29] J. Garrett, "Ajax: A new approach to web applications," blog posting, available online at http://www.adaptivepath.com/ideas/ajax-new-approach-web-applications, February 2005.

[30] J. Allaire, "Macromedia Flash MX—A next-generation rich client," *Macromedia White Paper*, pp. 1–2, 2002.

[31] R. Hanisch, A. Farris, E. Greisen, W. Pence, B. Schlesinger, P. Teuben, R. Thompson, and A. Warnock III, "Definition of the Flexible Image Transport System (FITS)," *Astronomy and Astrophysics*, vol. 376, no. 1, pp. 359–380, 2001.