Earth Sci Inform (2008) 1:35–42
DOI 10.1007/s12145-008-0004-5

METHODOLOGY ARTICLE

Navigating through SPASE to heliospheric and magnetospheric data

Jan Merka · Thomas W. Narock · Adam Szabo

Received: 3 August 2007 / Accepted: 18 January 2008 / Published online: 12 March 2008

Abstract Virtual observatories have been introduced by the astrophysics community as an environment connecting distributed data sources with a unified interface. The heliophysics community soon recognized that they faced a similar problem of many distributed data sets with varying amount of information about them and several discipline specific virtual observatories have been established. Two of them, the virtual heliospheric observatory (VHO) and the virtual magnetospheric observatory (VMO), share a common architecture design with development efforts oriented towards a structured data search. This paper describes the VHO/VMO middleware and its components from metadata preparation and processing to the user interface.

Keywords Middleware · SPASE metadata · VHO · Virtual observatory search · VMO

Introduction

Finding and retrieving space physics data is still a rather daunting task even when the data are publicly available on the Internet. The space physics data environment consists of thousands of relatively small and many large datasets and, in principle, a web search engine should find them. Unfortunately, the relevance scoring methods successful in a web environment, for example the Google's PageRank method, are not suitable for space physics data sets because most resources are singular observations that are not linked to other resources (King et al. 2008b).

Initially a group of astrophysicists coined a term 'virtual observatory' (VO) for an environment where one was able to examine any patch of sky using collected observations in all wavelength ranges by connecting resources on the Internet (Astronomy and Astrophysics in the New Millennium [Decadal Survey], National Academy of Science, http://www.nap.edu/books/0309070317/html, 2001). In October 2004 at the NASA/LWS workshop in Greenbelt, MD, space physicists and data engineers laid the foundations for similar data environment that provides uniform access to data and services for specific science communities (Bentley et al. 2005). Currently at least ten virtual observatories (VxOs) are under development within the space physics community (Harvey et al. 2008).

Two of these virtual observatories share a common architecture design and development efforts oriented towards a structured data search. They are the virtual heliospheric observatory (VHO) and the virtual magnetospheric observatory (VMO) focused on serving the heliospheric and magnetospheric research communities, respectively. The VMO is a multi-tiered environment composed of two peer observatories, one located at NASA Goddard Space Flight Center (NASA/GSFC) and the other at University of California, Los Angeles (UCLA), that divided up tasks in a complementary fashion similar to open source

Editorial handling: P. Fox

J. Merka (✉) · T. W. Narock
Goddard Earth Sciences and Technology Center,
University of Maryland, Baltimore, MD, USA
e-mail: jan.merka@nasa.gov

J. Merka · T. W. Narock · A. Szabo
Heliospheric Physics Laboratory,
NASA Goddard Space Flight Center,
Greenbelt, MD, USA

development methods (King et al. 2008a). While the NASA/GSFC VMO development is geared towards a relational search, the UCLA VMO is working on a Google-like word search. This paper describes the VHO/VMO implementation of structured search and related components.

The presented VHO/VMO observatories serve particular space physics communities, heliospheric and magnetospheric, and so this paper contains many terms and abbreviations that may be less familiar to other audiences. Thus for readers convenience, an Appendix summarizes abbreviations frequently used throughout this paper.

Method

Bentley et al. (2005) define a virtual observatory as follows:

> A virtual observatory is a suite of software applications on a set of computers that allows users to uniformly find, access, and use resources (data, software, document, and image products and services using these) from a collection of distributed products and service providers. A VO includes registries based on a metadata model, front–end applications, and connections to data providers.

Thus, a VO is a complex distributed environment with a goal to provide a single point of discovery of data and related resources. We have decided to design the VHO/VMO using a modular *small-box* approach introduced by the Virtual Solar Observatory (Bogart et al. 2002) where individual components are small both in size and number of features but perform them extremely well.

Modularity

Advantages of the small-box approach are numerous and affect both development and day-to-day operation of VHO/VMO. In particular, such components are easier, faster, and cheaper to develop and to maintain because they are less complex. That makes them quicker to understand, test, and debug. Lightweight components with well-defined functionality can be simply chained in different ways to fulfill complex tasks. As a simple example, a query-engine finds matching data granules and passes the results to a format conversion tool that returns the results in a user-specified data format. Note that with the modular approach, it is quite simple to replace a component with a newer version or an alternative implementation. Different teams can create these components, each using various tools and/or programming languages.

Linking modular components together requires that directly chained modules understand each other. This may not be a problem when the modules are developed by a single team but quickly becomes a significant issue when disparate teams contribute components. In the latter case, to exchange information with data providers, services or other VxOs, a standard communication format must be used. This is the reasoning behind ongoing efforts to develop a common query language for VxOs that will facilitate inter- and intra-VxO communication (Narock and King 2008). The VHO/VMO actively participate in the development of this query language and plan to employ it for communication with other VxOs, services and data providers. Currently however, no standard communication method has been implemented for the existing VHO/VMO components because (1) no suitable candidate has been found and (2) this communication will not be exposed outside of the VHO/VMO middleware rendering the use of a standard communication unnecessary.

Metadata driven data searches

An important aspect of the VHO/VMO small-box approach is that data granules (files) are not necessary for answering user queries because metadata are used instead. The metadata describe salient properties of available data sets and other resources using data dictionary developed by the Spase Physics Archive Search and Extract (SPASE) consortium (Harvey et al. 2004). The SPASE data model has been adopted and is co-developed by NASA virtual observatories as a standard for describing resources in order to facilitate data search and retrieval across the space physics data environment (Harvey et al. 2008).

The SPASE data model simplifies management of metadata by introducing resources that describe sets of information likely to be referenced several times from other descriptions. For example, a spacecraft would be described as an Observatory resource while its magnetometer would be an Instrument resource. The current version 1.2.0 of SPASE has the following resource types: Catalog, Display Data, Numerical Data, Granule, Instrument, Observatory, Person, Registry, Repository, Service. The most widely tested and used resources within the VxO environment are Display Data, Numerical Data, Granule, Instrument, Observatory and Person while the other resources will be extensively employed in the later stages of VxO

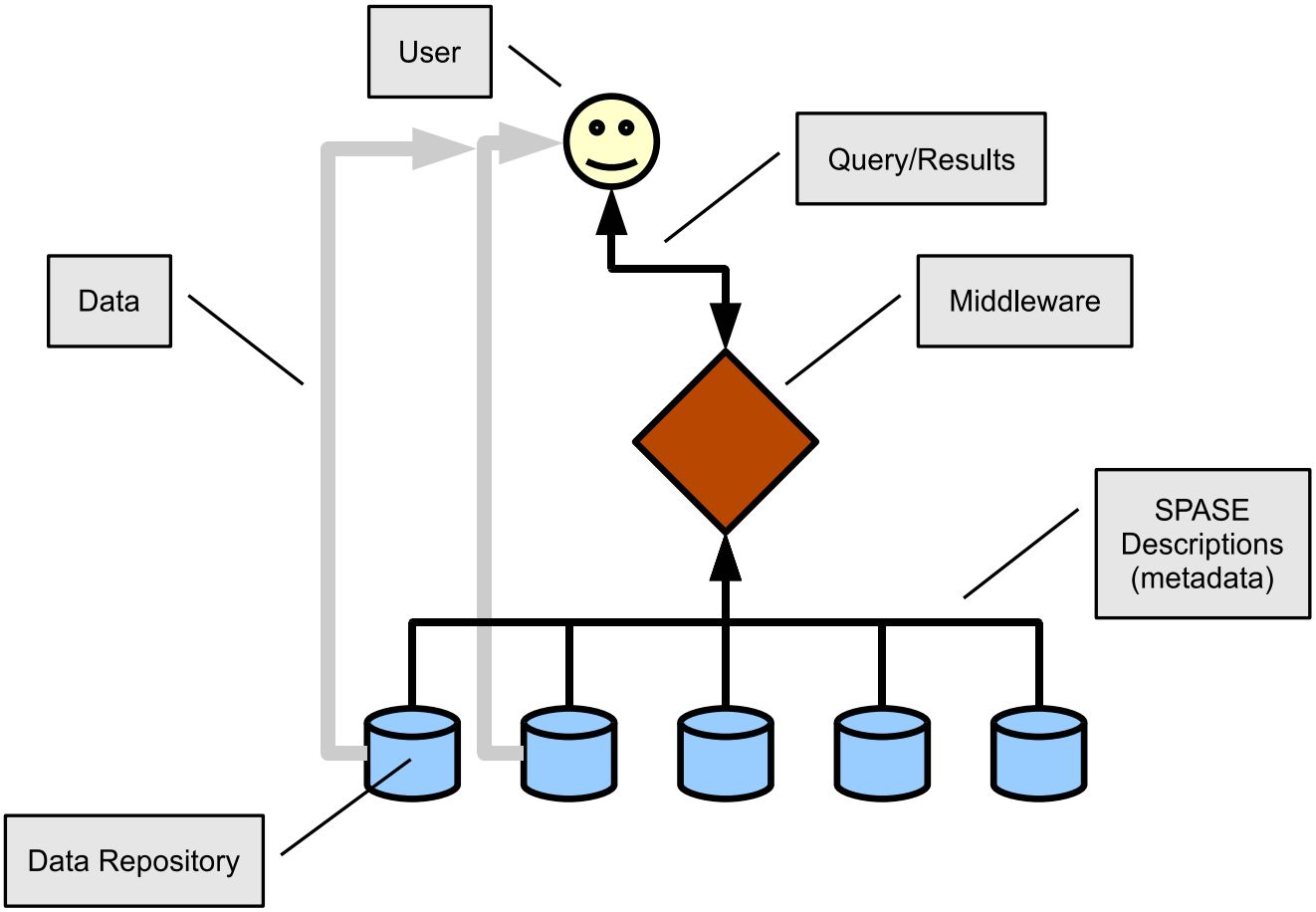

Fig. 1 Schema of the metadata and data flow within the VHO/VMO environment. SPASE descriptions (metadata) are collected by the middleware and used to answer user queries. The returned results include URLs of matching data granules that are directly obtained from data repositories bypassing the middleware

development, especially for introducing value added services (e.g., data reformatting, subsetting, plotting).

Figure 1 demonstrates information and data flows within the VHO and VMO environments. Data products residing in data repositories must be described in SPASE terms before they can be recognized by the VHO/VMO. Thus, the initial registration of a data product is a manual effort involving the exchange of product level metadata and dataset information. Subsequently, tools poll data repositories daily and harvest information regarding new and modified data. This information is also used to generate SPASE metadata for granule resources and populate VHO/VMO holdings. Although product level metadata are relatively static, a versioning system is maintained to log any changes and their history. User queries are answered by the VHO/VMO middleware with a list of matching data granules including URLs for direct download and associated SPASE descriptions. The descriptions give the user an opportunity to review the results before downloading the data directly from data repositories. Therefore, the direct data download eliminates a bandwidth bottleneck at the VHO/VMO servers because future services, e.g. data mining services, are anticipated to request large number of files per query.

The VHO/VMO middleware is an interconnected collection of components that keep track of metadata, process it and answers data queries via a web interface and, in the future, an Application Programming Interface (API), Fig. 2. All harvested metadata are stored in a registry before ingestion into a PostgreSQL database (Douglas and Douglas 2005). King et al. (2008a) describes in detail harvesting methodology that will be used by the VHO and VMO registries, and possibly across the entire NASA VxO environment.

In addition to SPASE metadata, the middleware needs also ephemeris (position, in particular) information about the data products to enable position-based searches. Traditionally, space physics data set access or search is time-based, the user must know the name of a data product and time interval of interest before submitting a query. The VHO/VMO aim for a query

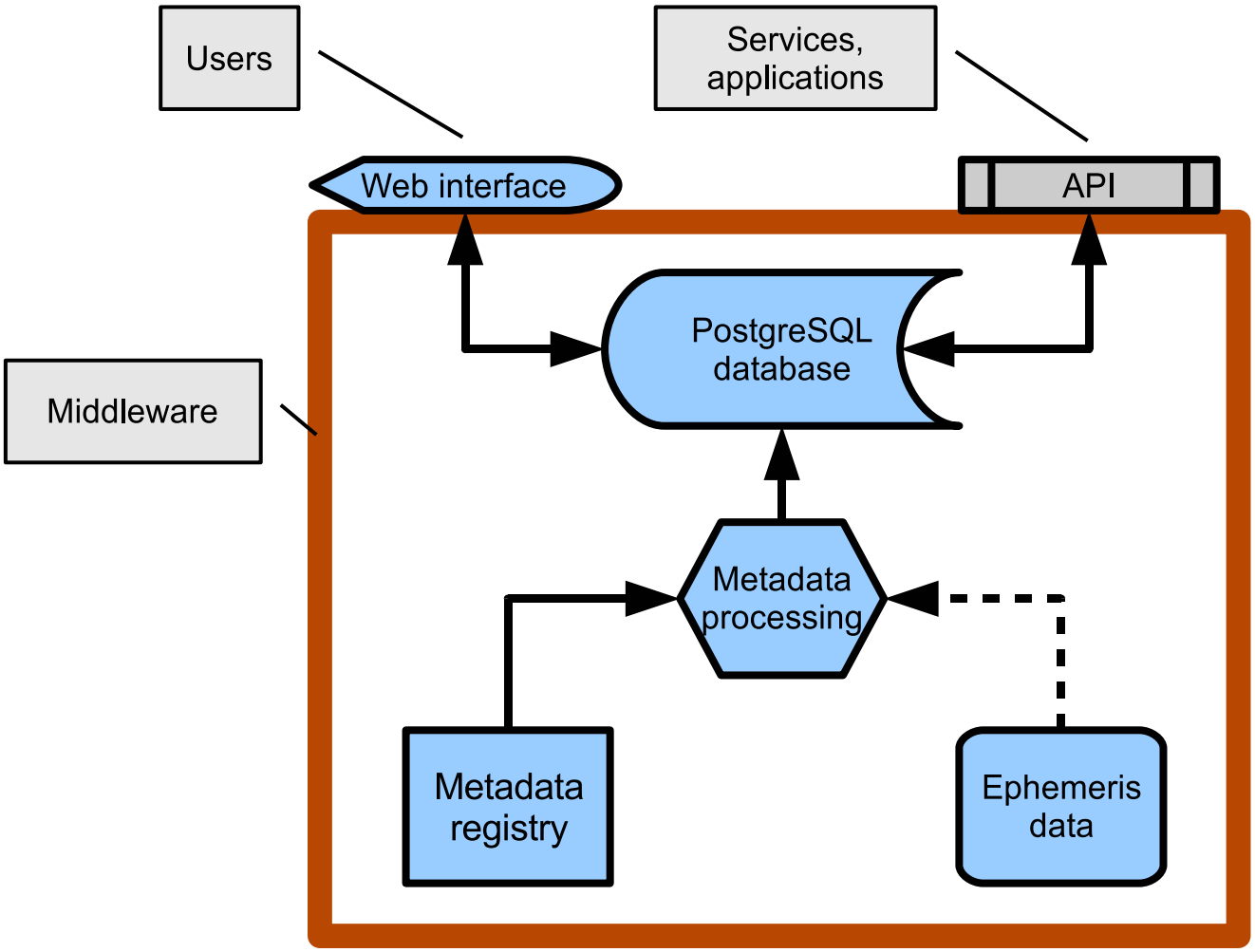

Fig. 2 Main components of VHO/VMO middleware. The middleware will provide two interfaces: A web interface for user access and an API (not implemented yet) to be used by services and client applications

engine with several basic search axes and their combinations. The current middleware prototype (version 0.8.5) handles time, position, measurement type (magnetic field, etc.) and parameter value types of searches. Therefore, a user can submit queries like 'What magnetospheric data is available at $X_{GSE} < -30\ R_E$?' without the need to identify products and time intervals first. (Here X_{GSE} stands for the first Cartesian component of a position vector in the Geocentric Solar Ecliptic (GSE) coordinate system and units are Earth Radii (R_E).) However, the position information necessary for such searches is generally not contained within the SPASE descriptions so a special component of the middleware collects ephemeris data and resamples them to a different time resolution with a configurable and optionally variable time step (Fig. 2). Note that we distinguish between remote-sensing and in-situ measurements when collecting position data: In-situ observations obtained by several instruments attached to a observatory are co-located and we can store only observatory positions. On the other hand, remote-sensing observations from instruments on the same observatory can observe completely different regions so we must link observed positions with the remote-sensing data products.

The configurable and variable sampling of positions significantly reduces number of records stored in the database and subsequently improves the search time. The time step is selected according to the observed position and, in the future, to the position rate of change in a given coordinated system. For example, in the case of observatories near the Earth's L1 Lagrange point (e.g., the ACE or WIND spacecraft), 1-h position sampling is more than sufficient while for a magnetospheric mission the 1-h time step may represent a substantial part of the orbit around Earth. Therefore, we are currently exploring several possibilities how to vary the sampling rate of position data stored in the database. A rather straightforward algorithm under consideration is to adjust the sampling rate as a function of radial

distance from the Earth to get a higher time resolution in the lower altitudes when spacecraft pass faster through different regions and a lower time resolution farther from the Earth.

Each discipline specific virtual observatory serves a different community that prefer to work with particular coordinate systems. Thus, the observed positions are converted and stored in coordinate systems that best serve the heliospheric and magnetospheric communities: The VHO offers Geocentric Solar Ecliptic (GSE), Geocentric Solar Magnetic (GSM) and Heliographic Inertial (HGI) coordinate systems while the VMO uses GSE, GSM and Solar Magnetic (SM) (Russell 1971). Positions in each coordinate system are kept in Cartesian, cylindrical and spherical representations. The VHO/VMO performs necessary coordinate conversions before storing the positions in the database to enable searching in any of the available coordinate systems and representations.

As mentioned above, the VHO/VMO middleware supports parameter value queries. In order to enable this feature we rely on the Extension element of the SPASE dictionary to append a VOTable (http://www.ivoa.net/Documents/latest/VOT.html) to Granule resource descriptions. The VOTable contain statistical properties of measured physical parameters: Average, median, minimum, maximum, standard deviation and availability. These statistical values are calculated over arbitrary time intervals when creating Granule descriptions. For example, the WIND Magnetic Field Investigation (MFI) data product uses 1-h long time intervals because the WIND spacecraft observes mostly near-Earth solar wind and such time intervals are sufficient for finding measurements pertinent to studied phenomena. Magnetospheric missions are currently sampled with 15-min resolution while variable time resolution will be tested in the future. We solicit community feedback to help us with finding the right balance between storing all data points in the database and a too coarse sampling.

If all data points were kept in the database, the middleware could, in principle, answer parameter value queries accurately but the amount of data is prohibitive (presently hundreds of Gigabytes growing past Terabytes soon). By using statistical values, the middleware can answer queries quickly and display most of the right results, thus significantly reducing the amount of data the researcher needs to download and analyze. In particular, the VHO/VMO middleware can answer questions like 'What data are available near Earth when the average value of the interplanetary magnetic field (IMF) Z-component in GSE coordinates is negative: $B_Z^{GSE} < 0\,nT$?' but cannot satisfy query 'What data are available near Earth when IMF $B_Z^{GSE} < 0\,nT$?'.

Data search capabilities and limitations

The PostgreSQL database is not exposed to a user directly but through either a web interface (http://vho.nasa.gov for the VHO or http://vmo.nasa.gov for the VMO) or, in the future, by an API. The API will utilize a VxO query language for information exchange over the Internet (Narock and King 2008). The web interface is written in PHP and Perl (http://www.php.net and http://www.perl.org) with a goal to provide an intuitive, clean and yet comprehensive interface to query submission and result presentation.

The actual data search is performed by the PostgreSQL database that holds the metadata (including the position information and statistical values of measured parameters). However, not all SPASE metadata content is entered into the PostgreSQL database, only what is needed for answering a query. The remaining metadata can be retrieved from registry based on SPASE ResourceID that is always part of the query response for every matching data granule. As a minimum, the response always carries a granule unique identification (Granule/ResourceID), download link (Granule/URL), and the time interval within the granule matching the query conditions so the user can download the data file, knows which part of the file is relevant and can obtain more information about the data using the ResourceID. In practice, the response is expanded by additional information (Granule/ParentID, Instrument/ResourceID, Observatory/ParentID, NumericalData/ResourceName, etc.) that helps the web interface to display the results in a comprehensive way.

Figure 3 presents a tree-like representation of the four main search axes that has been implemented in the web query interface. The main nodes are *Measurement type*, *Date/time*, *Position*, and *Parameter values*. Certain nodes provide children nodes that can be exposed if the user wishes. For example, the VMO web interface offers measurement types *Magnetic field* and *Ephemeris* and selecting one of these terms, say *Magnetic field*, would find all data products that provide magnetic field measurements while no prior knowledge of product, instrument nor observatory names is necessary. On the other hand, a more experienced user seeking magnetic field measurements from the WIND MFI instrument can expand the *Magnetic field* node and chose that particular instrument only. Figure 3 shows that the *Measurement type* axis spans the measurement type, observatory, instrument and product name but only

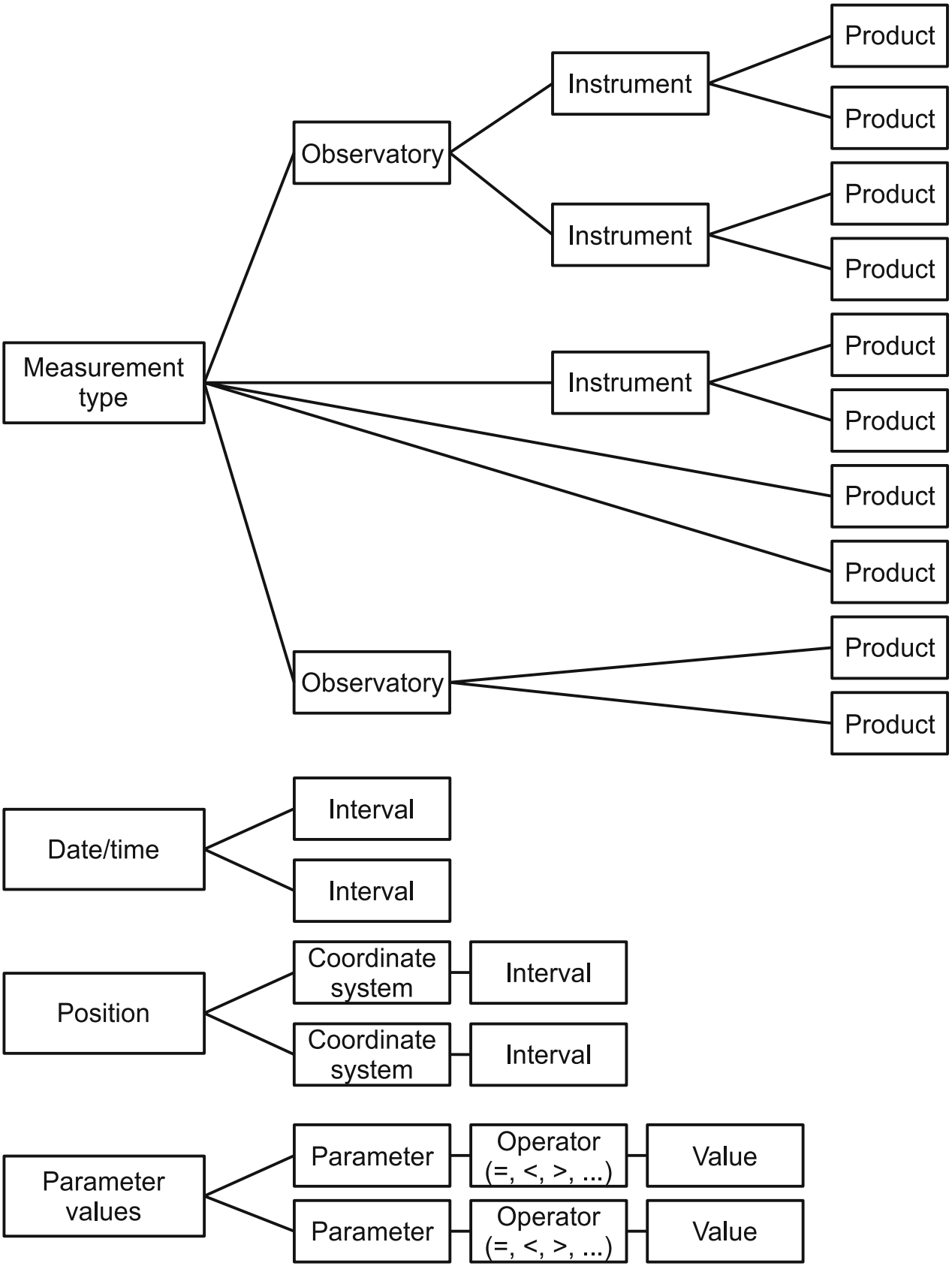

Fig. 3 Depiction of the four main search axes implemented in the VHO/VMO web query interface version 0.8.5. The 'Measurement type' branch exposes Observatory or Instrument nodes only when they have at least two children nodes. See the text for more detailed description

nodes with more than one child are displayed in order to shorten the branch path. Note that Fig. 3 displays only up to two children per node as an example while in fact the number of child nodes depends on the database content.

The three remaining search axes are shorter and require numerical values to completely define the query. In the *Date/time* branch, each leaf node defines a time interval while leaf nodes in the *Position* branch define a space volume in a particular coordinate system and representation. Leaf nodes on the *Parameter values* axis set the physical parameter, its statistical attribute (average, median, etc.), coordinate system if the parameter is a component of a vector, comparison operator (=, <, >, etc.) and a parameter value.

When multiple constraints are selected, an intersection of matches for each constraint is returned. In other words, all conditions must be satisfied at the same time to identify a matching granule. Results of query terms belonging to the same parent search node are, however, merged together as a union and then intersected with matches from other query terms. As a demonstration, we can select the following criteria:

1. Magnetic field data from the ACE spacecraft
2. Magnetic field data from the WIND spacecraft
3. Position at $X_{GSE} < -100\ R_E$
4. Position at $X_{GSE} > 200\ R_E$
5. Time between 1999-01-01 00:00:00 and 1999-01-10 00:00:00 UTC
6. Time between 2000-01-01 00:00:00 and 2000-01-01 00:00:00 UTC
7. IMF average $B_Z^{GSE} < 0\ nT$

Criteria 1 and 2 share the same parent node so they are grouped as a union (logical operator OR), the same is true for item pairs 3-4, and 5-6. The groups and item 7 do not have common parents so they must intersect (operator AND). The query submitted to the PostgreSQL database is then:

> ((Magnetic field data from the ACE spacecraft) OR (Magnetic field data from the WIND spacecraft)) AND ($X_{GSE} < -100\ R_E$ OR $X_{GSE} > 200\ R_E$) AND (1999-01-01 00:00:00 to 1999-01-10 00:00:00 OR 2000-01-01 00:00:00 to 2000-01-01 00:00:00) AND (average $B_Z^{GSE} < 0\ nT$)

It is apparent that the current web interface is rather restrictive because it does not allow arbitrary grouping of constraints nor selection of the logical operands. Thus it is not yet possible to construct queries like 'Find day side magnetosheath data from Geotail when WIND observed average IMF $B > 30\ nT$'. We are considering several different approaches that would enable this functionality on the VHO/VMO web interface. One of our primary goals is to provide an intuitive search interface that would require little or no learning from space physics researchers so we plan to provide a standard interface with reasonable default behavior as implemented in the current prototype, and an advanced interface where users will have full control over the query construction.

Discussion and conclusions

The architecture of the VHO/VMO middleware and its components is maturing rapidly with the first stable version (1.0) release planned for December 2007. All the major components are already in place (Fig. 2) except for the API which will be defined and implemented once the VxO query language is stabilized (Narock and King 2008). Nevertheless, significant amount of work will be spent on improving database performance and the web interface because they are both extremely important factors in user experience with the VHO/VMO. In order to attract and increase a user community, user feedback and feature requests are accommodated in the design.

The VHO/VMO are dependent on metadata and the SPASE data model in particular. We are actively contributing to the development of the SPASE data model in collaboration with other VxO teams. Ideally, the SPASE dictionary will be eventually comprehensive enough to capture properties of all heliophysics data products in a cross-discipline manner. But before we reach that point, many SPASE model versions will be released and it is likely that resource descriptions conforming to different version will coexist. This fact bears an important implication for the VHO/VMO and all other SPASE-speaking VxOs: The VxOs must be prepared to handle metadata produced according to different SPASE data model versions. A possible solution may be the use of XSLT stylesheets for transforming SPASE metadata from one version to another.

Acknowledgements This work was supported by the National Aeronautics and Space Administration under Grant No. NNX07AC95G issued through the virtual observatories for Solar and Space Physics Data (S^3CVO).

Appendix: List of Abbreviations

ACE	spacecraft (http://www.srl.caltech.edu/ACE)

API	application programming interface
GSE	geocentric solar ecliptic coordinate system
GSM	geocentric solar magnetic coordinate system
HGI	heliographic inertial coordinate system
IMF	interplanetary magnetic field
nT	nanotesla
R_E	earth radius
SM	solar magnetic coordinate system
SPASE	space physics archive search and extract (http://www.spase-group.org)
SQL	structured query language
VHO	virtual heliospheric observatory (http://vho.nasa.gov)
VMO	virtual magnetospheric observatory (http://vmo.nasa.gov)
VO (also VxO)	virtual observatory (for community "x")
WIND	spacecraft (http://pwg.gsfc.nasa.gov/wind/shtml)
XSLT	extensible style language transformation, the language used in XSL style sheets to transform XML documents into other XML documents

References

Bentley R, Bogart R, Davis A, Hurlburt N, Mukherjee J, Rezapkin V, Roberts DA, Szabo A, Weiss M (2005) A framework for space and solar physics virtual observatories. Available on the web at http://lwsde.gsfc.nasa.gov/VO_Framework_7_Jan_05.pdf

Bogart RS, Tian K, Hill F, Wampler S, Martens P, Davey A, Gurman JB, Dimitoglou G (2002) The Virtual Solar Observatory design proposal. http://virtualsolar.org/docs/VSO_strawman_20021125.pdf

Douglas K, Douglas S (2005) PostgreSQL. 2nd edn. Publisher Sams. (ISBN 978-0672327568. 1032 pages)

Harvey CC, Thieman JR, King T, Roberts DA (2004) SPASE—Space physics archive search and extract. In: Proceedings of ensuring the long term preservation and adding value to the scientific and technical data, pp. 5–7, October 2004, ESA/ESRIN WPP-232

Harvey CC, Gangloff M, King T, Perry CH, Roberts DA, Thieman JR (2008) Virtual observatories for space and solar physics research. Earth Sci Inform. doi:10.1007/s12145-008-0008-1

King T, Merka J, Walker R, Steven J, Narock T (2008a) The architecture of a multi-tiered virtual observatory—The VMO. Earth Sci Inform. doi:10.1007/s12145-008-0006-3

King T, Narock TW, Walker R, Merka J, Steven J (2008b) A brave new (virtual) world: distributed searches, relevance scoring and facets. Earth Sci Inform. doi:10.1007/s12145-008-0002-7

Narock TW, King T (2008) Developing a SPASE Query Language. Earth Sci Inform. doi:10.1007/s12145-008-0007-2

Russell CT (1971) Geophysical coordinate transformations. Cosm Electrodyn 2:184–196