Scientific Resource Access System: A Concept for Getting "Living With a Star" Information to Do Science

Rose A. Daley, Elisabeth A. Immer, Brand I. Fortner, and Michele B. Weiss

Answering complex science questions often means that a scientist must find and use a variety of scientific resources, including static and dynamically generated data in different formats, complex data assimilation models, and conversion and analysis tools. Although the Internet has led to an explosion of accessible scientific resources, it has also created a Babel of incompatible formats, data descriptions, and access methods from many disparate sources. An evolving prototype called the Scientific Resource Access System (SRAS) is being developed that uses metadata and commonly understood scientific concepts to provide easier discovery and access to these necessarily heterogeneous resources for "doing science." This concept simplifies the integration of distributed scientific resources within a single system and enables the interconnection of complex data systems. An additional goal of SRAS is to eliminate the need for specialized integration software and the need to force data standardization on participating resources. Both are infeasible in Internet-based scientific data environments, where the additional effort and expense required for such solutions would effectively eliminate a substantial quantity of important scientific resources. Instead, the SRAS uses existing protocols and formats along with robust metadata to simplify integration.

INTRODUCTION

The Scientific Resource Access System (SRAS) is being developed to support a NASA initiative called Living With a Star (LWS),[1] but it encompasses the entire solar-terrestrial scientific discipline. LWS is a space science exploration program covering the domain from the Sun all the way to the Earth (Fig. 1) for over a decade of missions, observations, and experiments. Therefore, LWS includes a wide variety of scientific disciplines (e.g., solar, space, ionospheric, and magnetospheric physics), a large amount of valuable legacy data, and an indeterminate variety of future data and data formats, all of which must be combined to support broad, interdisciplinary approaches to scientific analysis.

The fundamental goal of the SRAS is to give the space science community a simple and complete method to gather and share static and dynamically generated data, metadata, documentation, and model results, as well as associated analysis tools and services, without

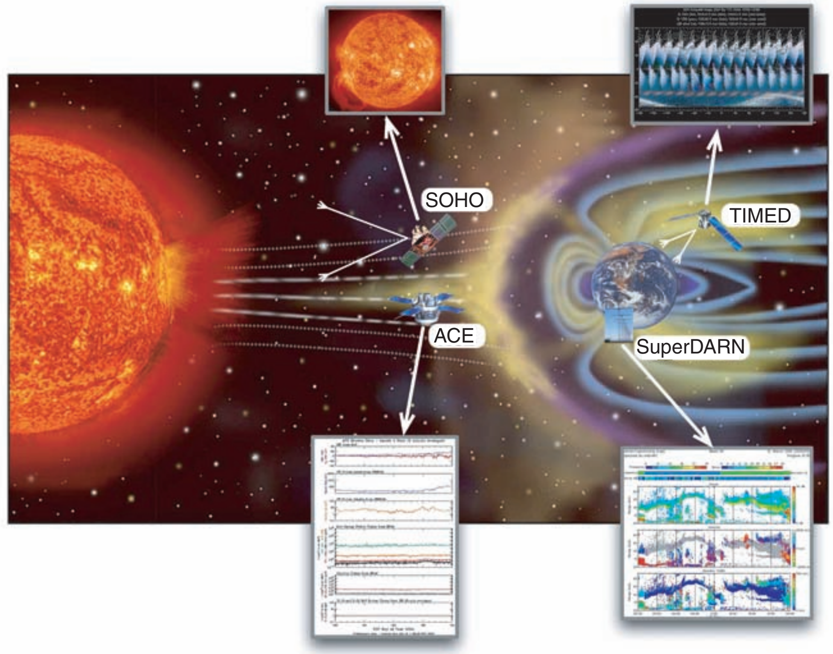

Figure 1. The solar-terrestrial science domain. Living With a Star science encompasses myriad scientific disciplines, scientific resources, and data formats. The ACE, SOHO, TIMED, and SuperDARN spacecraft illustrate the diversity of resource types needed by the solar-terrestrial scientist.

The heart of the SRAS is a conceptual model of commonly understood scientific concepts, augmented by robust metadata supplied by the creator of the resource to facilitate resource discovery and access. The SRAS does not presume to perform all of the operations needed by solar-terrestrial scientists. Instead, it focuses on leveraging the capabilities developed within the space science and other research communities, and it provides unique capabilities

- To discover available resources for a "concept of interest"
- To discover relationships among resources
- To access heterogeneous resources from a variety of sources
- To simplify integration with other systems to augment capabilities

requiring all of these resources to be centrally located or administered. It acts as an intermediary between the user and the distributed data repositories, enabling access to known resources regardless of their format, geographic location, status (active vs. historical), or origin (measured vs. modeled). Furthermore, the SRAS simplifies the integration of resources by enabling the interconnection of complex data systems without the need to create specialized software for each system or resource. Finally, because of the long life cycle of the LWS initiative, the SRAS also anticipates frequent opportunities for technology transition as new technologies emerge and their viability is demonstrated, both to improve its capabilities and to replace obsolete technology. Specific considerations for easing this transition, both into the SRAS as new technologies mature and out of the SRAS as technologies are demonstrated to be effective, have been incorporated into the system's design.

As the set of resources needed to perform the LWS program and solar-terrestrial science evolves and becomes increasingly complex, automated support will also become increasingly common. To effectively interact in this environment, the SRAS prototype includes an innovative Web-based user interface specifically designed for scientists familiar with the available resources as well as interdisciplinary scientists who may be less comfortable with those same resources. It also includes a system interface specifically designed so that other software systems can use the services provided by the SRAS.

THE CHALLENGES OF SOLAR-TERRESTRIAL SCIENCE

Consider the following scenario involving the challenges faced by the solar-terrestrial scientist:

> The winds and composition of the Earth's ionosphere are affected by geomagnetic storms and energetic particle events,[2] which are both produced by solar variability. Over time, which aspect of the solar variability dominates the ionospheric properties?
>
> A scientist could use any number of resources to investigate this issue. The Advanced Composition Explorer (ACE) spacecraft measures solar wind properties that drive geomagnetic storms; the Global Geospace Science Program (Wind) spacecraft is a backup when data from ACE are not available. Solar energetic particle measurements also come from ACE, while measures of fluxes in the magnetosphere come from the Fast Auroral SnapshoT (FAST), Solar Anomalous and Magnetospheric Particle Explorer (SAMPEX), Geostationary Operational Environmental Satellite (GOES), Polar Orbiting Environmental Satellites (POES), and Defense Meteorological Satellite Program (DMSP) spacecraft. Data on the evolution of geomagnetic storms can be derived from the Imager for Magnetosphere-to-Aurora (IMAGE) spacecraft's Far Ultraviolet

Imager (FUV) instrument, and the IMAGE/High Energy Neutral Atom (HENA) imager measures remote sensing of the aurora and plasmasphere, with the POLAR spacecraft's Ultraviolet Imager (UVI) serving as backup. The scientist can also get data on ionospheric properties from the Global Ultraviolet Imager (GUVI), TIMED Doppler Interferometer (TIDI), and Sounding of the Atmosphere using Broadband Emission Radiometry (SABER) instruments on the Thermosphere, Ionosphere, Mesosphere Energetics and Dynamics (TIMED) spacecraft.

A significant challenge for the scientist is data selection, which depends on the locations of the spacecraft in their orbits to ensure that they are measuring the same places at the same times. If lists of storms and particle events exist, they can help to identify the places (time intervals) to start looking. Otherwise, data for the entire time interval must be searched to develop a list. Once an event time is known, does the scientist have to pick up all the data to decide if the observing situation is good? Or is there a way to understand what the next limiting factor is? The scientist must also choose an equal number of quiet times to ensure that the changes measured in the ionosphere are not just part of the natural variation. With several hundred events occuring per solar cycle, this task requires a lot of effort just to get to the data of interest.

To address these problems in the current data environment, the scientist would need to find each appropriate Web site, visit each individual instrument's or mission's Web site, find and retrieve the information for a specific resource from that instrument, parse the resource manually, and then manually merge the information. The difficulty here is not that the resources aren't already available but that each individual resource is available only from its own site or facility, which may provide only a limited search capability. Thus the scientist must be familiar with all of the Web sites and other access mechanisms that supply resources of interest. Every NASA mission now has one or more Web sites, although they often confine data distribution to downloading existing displays or low-level data for use with custom software. While many standards and storage formats have been developed, each is optimized for its particular scientific discipline or to address specific mission constraints, and none of these standards and formats have been widely adopted or supported by low-cost commercial software.

Because of this heterogeneous environment (notionally depicted in Fig. 2), undue effort is diverted from scientific investigations to understanding and overcoming each particular resource's specific acquisition circumstances. Cross-mission studies and investigations by scientists outside of the missions are often precluded simply because the effort to acquire data is so intimidating. To effectively perform solar-terrestrial integrated science, more capable systems that can provide unified access to multiple, physically distributed resources, available in different formats and at various levels of support, are needed. The SRAS is being designed and prototyped specifically to provide a means for solving these problems and making the solution available to a broad community of space scientists.

SRAS OVERVIEW

Concept

The SRAS concept is driven by four primary constraints: a long-term life cycle, integration of heterogeneous resources, involvement of a diversity of scientists, and finite science budgets.

1. Long-term life cycle for both development and deployment. Because LWS satellites will not be launched for several years, many challenges will not be fully experienced for quite some time. Given the rapidly changing

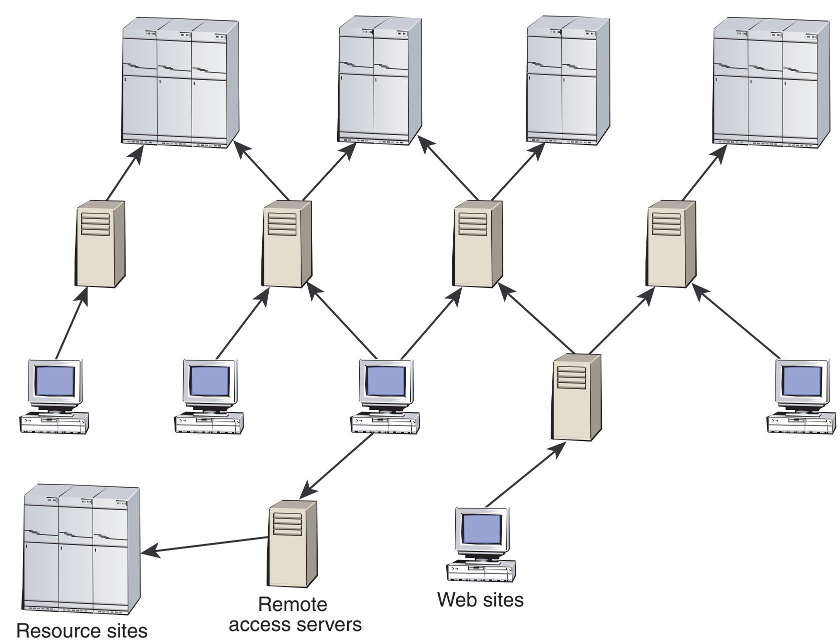

Figure 2. The LWS data environment. Scientific resources used by solar-terrestrial scientists are highly heterogeneous, often supporting multiple access mechanisms, including other scientific resources.

information technology field of recent years, contemporary standards will almost certainly be considerably altered or even replaced within this timeframe. In addition, new technologies will emerge and existing technologies will become obsolete.

2. Integration of diverse, heterogeneous resources. Many relevant scientific resources already exist (although some are in forms that will be expensive to update), new resources will be developed within the LWS program, and new resources important to solar-terrestrial scientists will be developed independently. Again, these resources will include static data sets as well as tools and models that generate new data dynamically. Furthermore, new resources will continually be developed even as the system evolves. It is critical that these resources be used cost-effectively.

3. Involvement of a diversity of scientists. Questions and problems that solar-terrestrial scientists raise will involve investigating a broad range of complex and interconnected problems. These will necessitate interdisciplinary use of the available scientific resources and flexibility in providing information for solar-terrestrial scientists. A single user interface or technology is unlikely to adequately satisfy these needs. Scientific resources will need to be accessed by, and supplied to, a variety of user interfaces as well as other systems. In addition, more scientists will be using resources in domains other than their primary area of expertise and familiarity; experts in some disciplines will be neophytes in others.

4. Finite science budgets. Scientific budgets are finite, and the primary emphasis must be on scientific pursuits. The ability to integrate new resources, user interfaces, and other systems must not require expensive new development, conversion, or translation of equivalent information, formats, and protocols.

APL's approach to dealing with these constraints has focused SRAS development in two areas: (1) minimizing the impact of technology and requirements changes and (2) minimizing the cost of supplementing and using SRAS information at the boundaries (user and system interfaces) and the cost of resource site interaction. Since the SRAS prototype is being implemented in stages, it will have progressive releases to the community to gain feedback. The SRAS is designed to be flexible enough for individual software components and services to be readily replaced, modified, or added. Subsequent releases will have (often radically) altered components as new capabilities are added to address specific usability feature requests from users or to incorporate new technologies as information technology and community protocols and standards evolve. In particular, a number of research efforts currently under way at APL and in various research communities are attempting to overcome the diversity of terminology, metadata, and file formats and standards that currently impedes the use of available scientific resources. The SRAS is building on, integrating with, incorporating the results of, and in some cases transitioning technology to these ongoing research efforts.

The heart of the SRAS is a conceptual model (described in more detail later) that captures the primary concepts employed by the SRAS and its users, both scientists and resource providers. This descriptive and intuitive model of the solar-terrestrial problem domain is critical to providing enough overall understanding to space scientists so that they can work comfortably in the domain and to computer scientists so that they can build systems to support the science missions. In particular, a significant amount of abstraction, including multiple levels of abstraction, is needed to successfully encapsulate the very broad nature of solar-terrestral science (multidiscipline, multimission, and multitechnology) in support of a diverse user community. The SRAS model will insulate the software from myriad sets of terminology and standards that are a fact of life in any such community, while at the same time providing support for them at minimal cost. Most importantly, this model enables "smarter" search capabilities beyond keyword matching, which is critical for interdisciplinary studies.

The SRAS is a distributed system that accesses and is accessed by other (also distributed) systems. However, while it is designed for its software and database to be distributed, specialized SRAS software will not have to be installed at resource provider sites; instead, it will use existing software at the site. The SRAS capitalizes on existing standards for remote access protocols and uses robust metadata to provide the details required by the access protocol. This feature is key to maintaining future SRAS adaptability. It also provides a capability to use legacy or heritage data sets where funding may not be available to modernize existing systems and formats. In some cases, this approach will limit the capability of the SRAS to that supplied by the resource provider and the remote access protocol. Finally, the SRAS includes an innovative user interface (described later), but in addition is intended to support other externally developed user interfaces as the science needs of the LWS program evolve over time.

The provider of scientific resources registers the resources with the SRAS. The motivation is to support the provider's activities by enlisting a larger community to use those resources. Part of the registration process includes robust metadata that, in addition to providing information for accessing the site, also gives descriptive information about the resources available. This information can include descriptions of methods used, errors, and assumptions, as well as references to papers and contacts that can provide more information. These resources are then mapped into the conceptual model to enable searching and access. Requests for information are first translated into conceptual model concepts; matching resources are subsequently found, and specific requests for the resources are generated. Because of the diversity of scientific inquiry, the results of these specific requests are not fused into a single coherent answer but are supplied as a set or collection of discrete products. While a future capability is envisioned to chain together requests and to provide pre- or post-processing of the data (e.g., format translation, if a translator is available), including fusion of the data products, these capabilities are not currently being addressed.

Structure

The SRAS is a Java Web application, and its processing is designed to be distributed, both for scalability and flexibility. Several standards and current technologies already exist, and the SRAS takes advantage of these as much as possible, primarily to maximize flexibility. To isolate operations for presenting information to users from the background support processing, the SRAS uses Java Server Pages (JSP) and Java Servlet technology. In the near future, we expect to formalize the system interface to the background processing, using other modern technologies for distributed computing—Web services and possibly Common Object Request Broker Architecture (CORBA) technologies. As integration with other systems progresses and the information content for these operations becomes more clearly defined, the interface will also mature. Internally, the SRAS uses relational database technology (Oracle) for storing persistent data (e.g., metadata and interim request results), which are accessed using the JDBC database connection protocol to provide remote access capability. We expect to incorporate metadata representation using XML and other technologies as they mature, particularly those that are part of the Semantic Web, a standard framework that allows data to be shared and reused more easily.

The SRAS software is logically organized into seven components (Fig. 3), which are groupings of related operations, not formal software components using component technology. This partitioning was driven by two concerns: (1) the desire to encapsulate technologies to

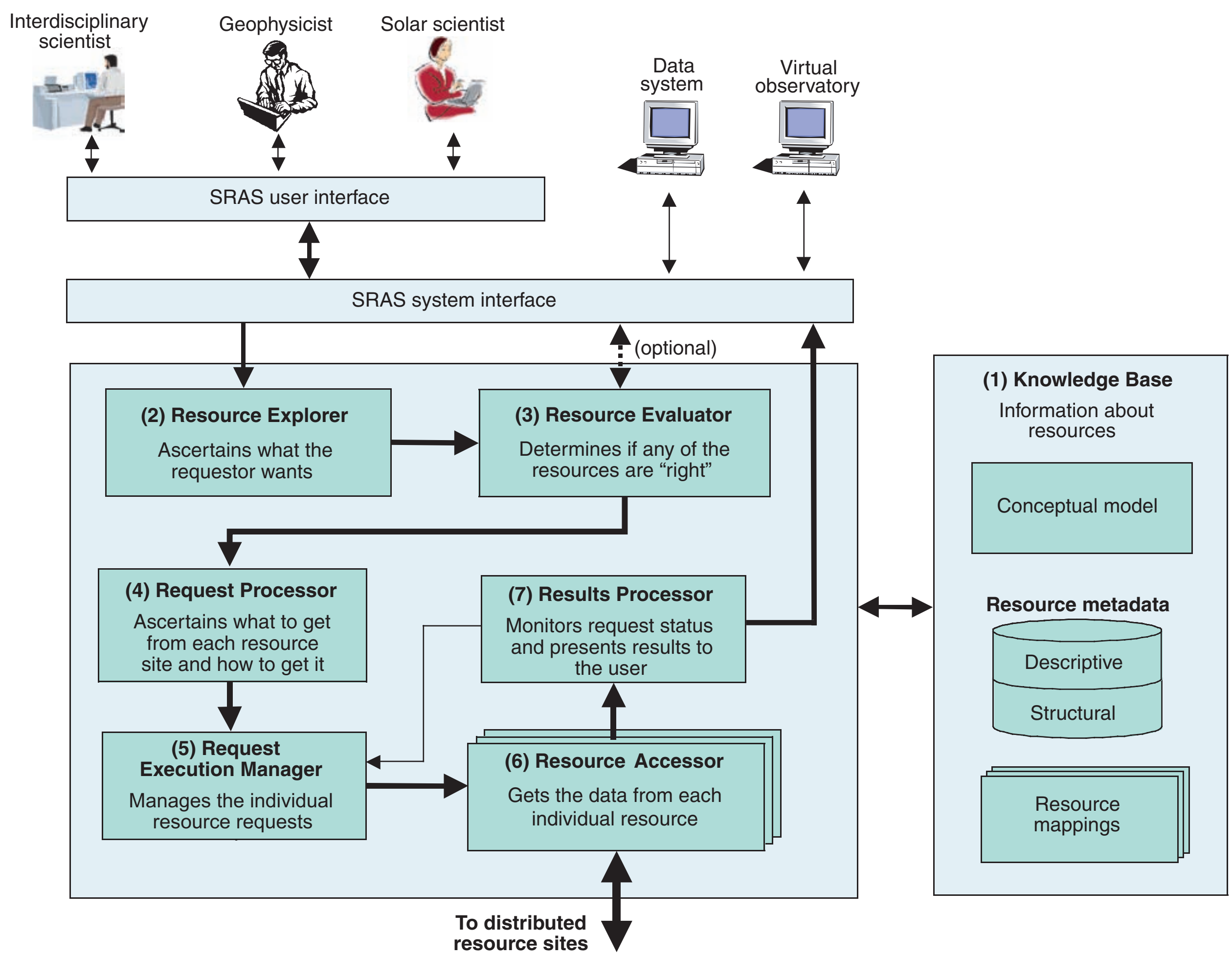

Figure 3. The SRAS structure. The system is designed for flexibility to facilitate incorporation of new features and resilience as technology evolves.

minimize perturbations due to technology evolution and transition, and (2) the need to distribute, replicate, and load-balance multiple instances of components to eliminate bottlenecks and improve performance. A high-level description of each component and its envisioned evolution capability follows.

1. Knowledge Base (KB). The KB houses the conceptual model, resource mappings, and locally stored metadata. It is used by all other SRAS components to support their processing. The entire KB is currently housed in a relational database, but the metadata are expected to be provided in XML and may be stored that way in the future. The conceptual model and mappings will incorporate ontologies and Semantic Web technology as these become more mature. It is assumed that metadata will be distributed at some point, although some will probably remain local, particularly for heritage data.

2. Resource Explorer (REx). The REx provides the ability to elicit and refine a user request, directly supporting the user interface. An important innovation enabled by the REx, in conjunction with the KB, is the ability to respond to user requests that do not require deep knowledge of a particular discipline, such as spacecraft names, instruments, or even spatial coordinates.

3. Resource Evaluator (REv). The REv determines which available resources "match" a specific user request. This evaluation is partially automated and partially interactive. It goes beyond keyword matching to information in the metadata (as typically performed by search engines such as Google), instead using relationships in the conceptual model to identify relevant resources that may not have matching keywords. APL has done some experimentation with further automation by adding properties to the relationships in the conceptual model that the user can select via a "Helper Model" for specific situations (e.g., high-velocity particle propagation from the Sun to the Earth). Work to incorporate dynamic calculation of relationship properties ("virtual metadata") is also under way to make these Helper Models more responsive to the specific circumstances in the user request and to enable the ability to plug in new models. The REv is expected to evolve as Semantic Web technologies mature, enabling the use of semantic information to determine matches. The REv will remain partially interactive to allow scientists to manually refine searches as their particular needs dictate. Thus, like the REx, the REv must directly support a user interface for user interaction. In addition, where resource providers have preview or remote examination capability, the REv will allow scientists to "try before they buy."

4. Request Processor (RQP). The RQP translates conceptual requests created using the conceptual model into individual, specific requests for each resource. The RQP relies heavily on QUICK technology (a universal interface with conceptual knowledge), which is an automated query formulation system developed at APL.[3] QUICK exploits conceptual-level database design knowledge (which is not the same as the SRAS conceptual model that is part of the KB) to generate SQL queries from high-level data requests. For the SRAS, this technology is being augmented to support the additional formats and access protocols present in the space science domain.

5. Request Execution Manager (REM). The REM organizes specific resource requests into a logical progression. For static, synchronous requests such as SQL queries or SOAP requests, the REM is not strictly needed. However, as more dynamic request capabilities are developed, such as virtual metadata and multistage requests (e.g., automatic format conversion), the REM becomes increasingly important. Also, as asynchronous requests (e.g., those that invoke data assimilation models that can take hours to run) become more common, the REM will need to coordinate with the Results Processor to maintain continuity and track status for the user.

6. Resource Accessor (RA). The RA is key to eliminating the need to develop specialized software for each new resource. In the SRAS, there are multiple RAs, one for each type of remote access protocol supported by the system. Each type of protocol has been parameterized, with information unique to the resource (e.g., its URL) specified in the metadata, which are referenced by the RA to perform the operations to access the resource itself. There are limitations to this technique, particularly for older protocols such as FTP, since it relies on a predictable mechanism (e.g., consistent naming conventions) for relating metadata information to individual resources. Currently, the SRAS supports the FTP (although this is limited by the site's capabilities, particularly when a regular naming convention is not used), Common Gateway Interface (CGI), and JDBC protocols. In the near future, RAs for Web services and resource-specified application programming interfaces will be added.

7. Results Processor (RSP). The final component is the RSP, which is responsible for aggregating results and presenting them to the user. Again, this is a partially interactive process, allowing the user to view or plot certain types of data before downloading. This capability will expand as more tools, particularly analysis tools, and the capability to perform multiple operations in sequence are developed. In general, the SRAS will not develop new post-processing or fusion capabilities, but will take advantage of those available as scientific resources.

The User Interface: Enabling the Scientist

Again, a primary goal of the SRAS is to make it easier for scientists to find and access scientific resources. An important facet of such access is an easy-to-use and intuitive user interface to the sophisticated underlying technologies. One way to make an interface easy to use is to make it familiar. With that in mind, a representational survey of space science access portals was undertaken to experience the current state of the art, to see if there were any common themes in these portals, and to ascertain whether the prevailing designs were flexible enough to provide the interaction desired. Although powerful and often relatively easy to use, none of the surveyed portals provided an intuitive, physically based, homogeneous access to heterogeneous data. That said, the survey did highlight numerous features (e.g., a data availability plot and ways of specifying temporal ranges) that are very useful for scientists. Portals explored in some depth included the TIMED instruments noted previously,[4] the TIMED Mission Data Center,[5] the Virtual Solar Observatory (design),[6] National Space Science Data Center (NSSDC) ModelWeb,[7] NSSDC BowShock Browser,[8] University Partnering for Operational Support (UPOS),[9] Satellite Situation Center Web (SSCWeb),[10] Coordinated Data Analysis Web (CDAWeb),[11] National Oceanographic and Atmospheric Association (NOAA) Space Environment Center (SEC),[12] Coupling, Energetics and Dynamics of Atmospheric Regions Web (CEDARWeb),[13] and various Earth Observing System (EOS) sites.[14]

The SRAS is targeted to interdisciplinary researchers who may not be intimately familiar with the various useful sources of current and heritage data. This implies that the interface needs to satisfy the following requirements:

- The user must be able to transparently search across numerous resources, including models, data repositories, databases, and mission archives.
- The search must be based on physical parameters (spectra, particles, etc.), and not on satellites and instruments or even scientific disciplines.

These requirements form the core of the SRAS user interface design and the core of much of the SRAS design in general. For example, searching by physical parameter has profound implications for the SRAS conceptual model and for how resource providers are initially linked with SRAS. However, while the focus is on the interdisciplinary researcher who may not know resource details, this interface must also support users who *are* familiar with particular resources. In other words, search capability must be provided both by physical parameters *and* by resource, and in a way that is easy to use and intuitive for both user classes.

One problem with physically based searches is the probability that the user may get hundreds or thousands of results. A successful SRAS interface must make it easy to manage and weed out large result sets. Therefore, a vital component of any successful SRAS interface is its temporal, spatial, and spectral selection tools, needed for any search, either by physical parameter or by resource. Since these components will be by far the most heavily used parts of the SRAS interface, a well-thought-out, orthogonal, intuitive, and graphical set of temporal, spatial, and spectral tools is crucial to creating a positive and productive SRAS user experience. Therefore, the following requirements have been added:

- Restrictions (e.g., indices or parameter values) must be an integral part of the search process.
- The search process must be interactive and iterative so that the user can easily modify search parameters and quickly see a reduction in the results set based on an evaluation of the available resources. Summary descriptive information must be readily available during this process to aid in the evaluation.
- Since searches by necessity will be fairly complex, management of the search set (e.g., saved searches) must be part of the interface.

With these design goals in mind, APL developed a concept for a new kind of user interface that includes two ways of accessing resources: (1) *Browse*, for users who are aware of particular resources and just want quick access, and (2) *Query*, for users who want to do an interdisciplinary search using physical parameters. The space science domain was also arbitrarily divided into four regions, following the NOAA SEC portal (*Solar*, *Interplanetary*, *Geomagnetic*, and *Near-Earth)*, to help reduce the potential flood of resources returned for general requests (Fig. 4). All scientists understand this division even though the boundaries between the regions are fuzzy (and users can always bypass this division by selecting *All* for "Environment"). Some of this interface remains conceptual; many of the features of the *Query* conceptual interface have been implemented, for example, but the user interface implementation is still evolving as of this writing.

An important innovation for this interface was to divide the large number of search selections into only three broad categories: *Coverages*, *Parameters*, and *Restrictions*, for both Browse and Query searches:

- *Coverages*. Here the user selects the spatial and temporal range of the requested resources, independent of format or instrument. The user can also narrow the search to selected resource providers.
- *Parameters*. For Browse searches, the parameter interface looks similar to all other science data portals in that the user is presented with a list of available resources sorted by name. For Query searches, the parameter interface concentrates on physical parameters such as particle flux, magnetic field strength, spectral band, hydrodynamic parameters, etc.
- *Restrictions*. This category offers a way to reduce the number of matching resources by providing restrictions such as file format, involvement in particular events (e.g., a coronal mass ejection), or even the

value of a particular parameter (e.g., "only include resources when the geomagnetic index Kp < 3").

The user can create as many Coverages, Parameters, and Restrictions as desired, including mixing and matching items (e.g., making a Restriction apply to one Coverage but not another). Since these searches can get complicated, the ability to save, retrieve, and modify searches or search items (including those contributed by other users) from a saved history is an integral part of the interface.

Figure 4 shows a prototype interface. The left side of the screen summarizes the various Coverages, Parameters, and Restrictions the user has added (one of each, in this example), and which category applies to which (supported by the REx component). For simplicity, entries in each category are designated as C1, C2, . . .; P1, P2, . . .; and R1, R2, ... The right side of the screen shows a representative selection for a particular restriction. In this case, the user can restrict a search by environment, by a particular event, or by the values of two geophysical parameters. Note the item on the left side of the screen called Helper Model. This feature provides a powerful way of generating a complex search set by using a simple computational model to generate search parameters. For example, a single coronal mass ejection involves solar images, particle counts, auroral measurements, etc., at times that differ by as much as several days. A computational model of that particular event can automatically create the requisite Coverages, Parameters, and Restrictions, in the proper order. This part of the user interface is supported by the REx component.

After the search is executed, the SRAS presents a list of matching resources, as shown on the left side of Fig. 5. (Not shown is how the user can choose subselections of these resources for further processing.) On the right side of the screen, a rough draft of temporal and spatial coverage maps for the selected resources is displayed. The temporal coverage map (upper right of the screen) lists the resources on the vertical axis and the time range on the horizontal axis. The temporal coverage for each resource is shown as a horizontal line spanning the time range available for the resource selected. The spatial coverage map (lower right of the screen) is simply divided into four quadrants, one for each region (Solar, Interplanetary, Geomagnetic, and Near-Earth).

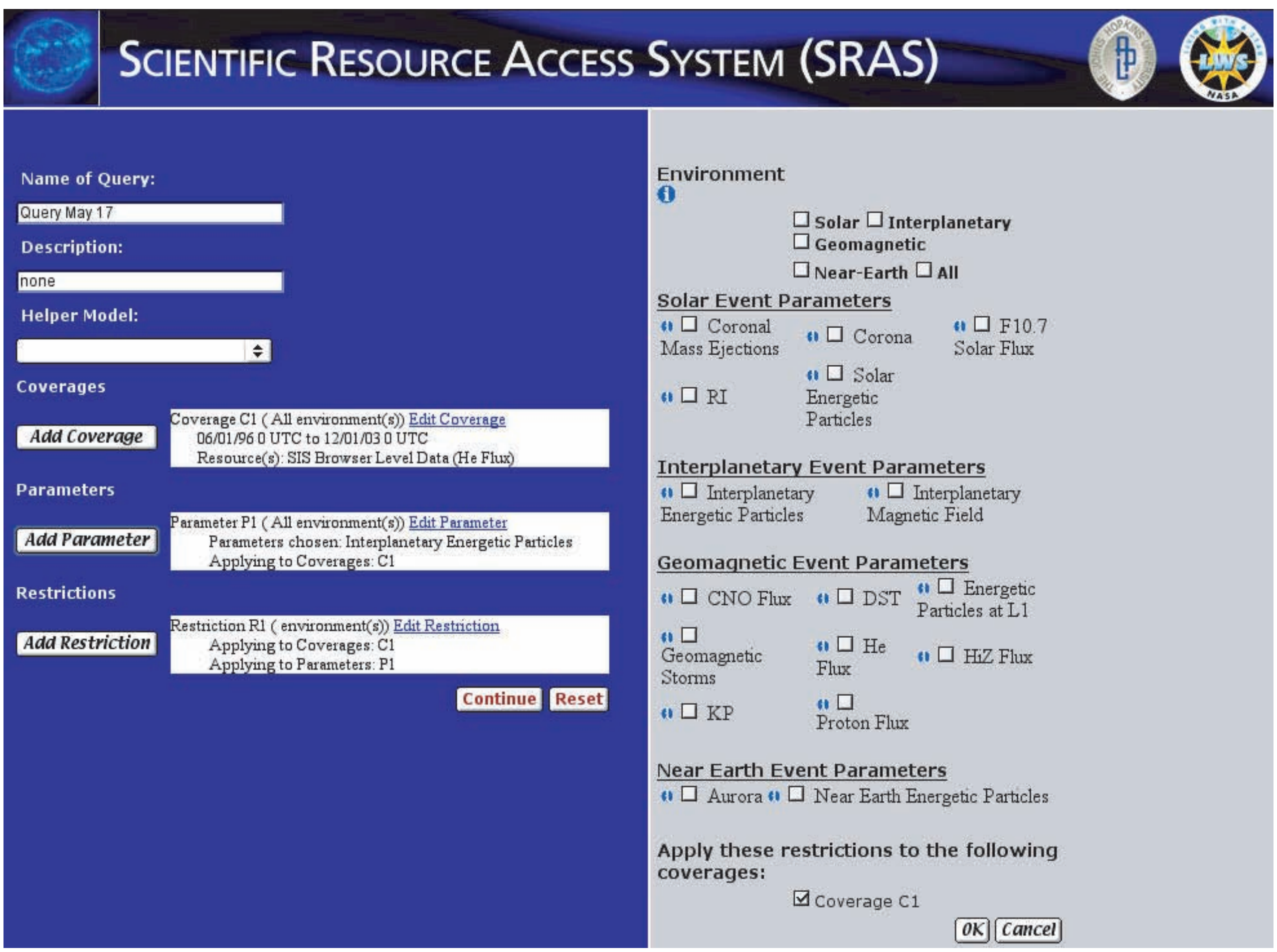

Figure 4. Requesting scientific resources. The SRAS user interface presents an intuitive organization that simplifies access for interdisciplinary researchers.

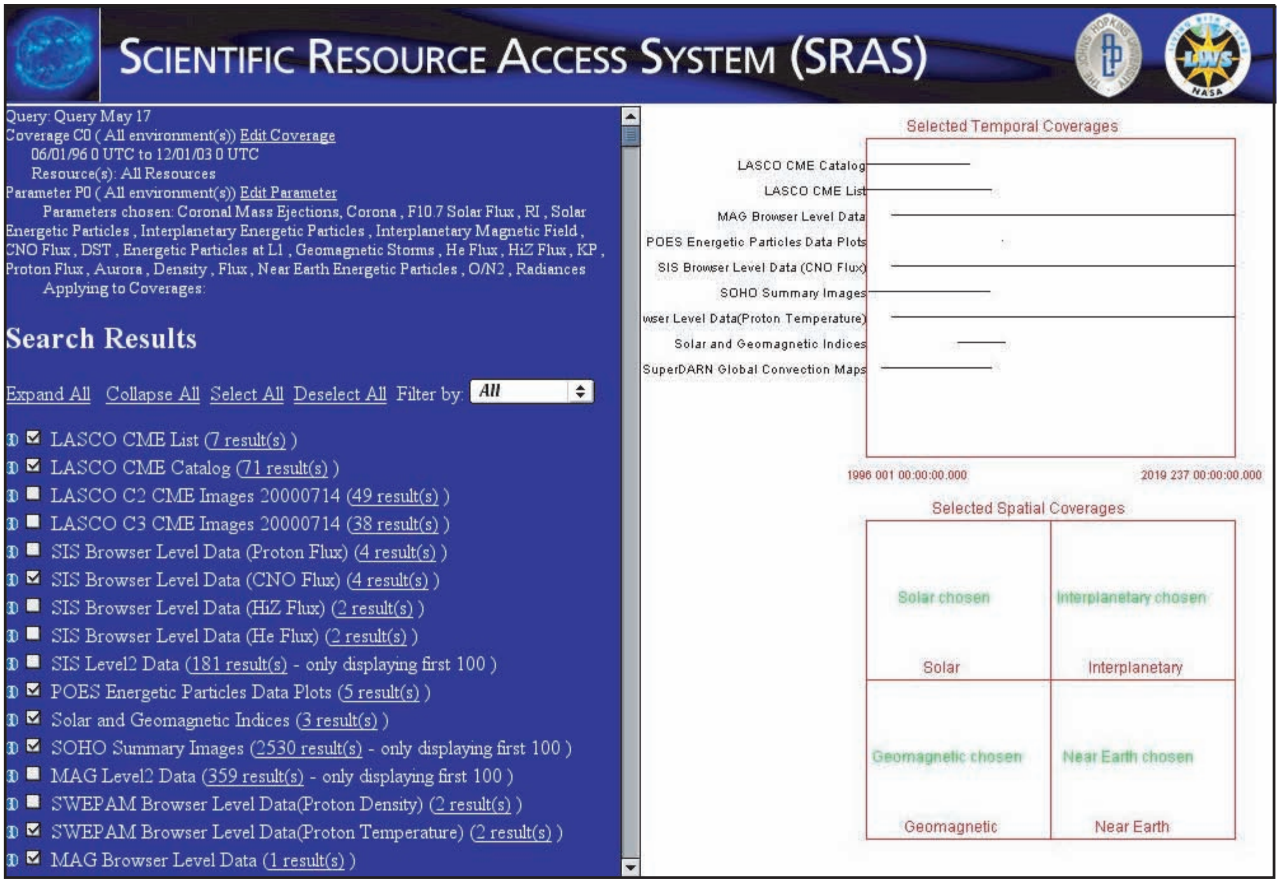

Figure 5. Selecting scientific resources. Temporal and spatial coverage maps make it easier for scientists to find the most useful resources without having to download them first.

When a resource is selected, a message indicating the region selected is displayed in the appropriate quadrant. The concept for the spatial coverage map includes a graphical indication of the actual spatial coverage for each resource, but this feature has not been implemented as yet. This temporal/spatial coverage map will be integral to the search process and will provide graphical coverage selection, resource availability, particular resource coverage, and limited visualization capability, such as showing parameter values in their proper temporal locations. The REv component provides the background support for these capabilities.

Not shown, but part of the current SRAS prototype, are screens to select specific resources to download, to visualize images, or to plot certain types of data, as well as to actually download the resources themselves.

Much work remains to expand and test the parameter lists and to provide the powerful, graphical, easy-to-use, and interactive temporal/spatial coverage map just described. Additional capabilities are still being developed, including user profiles, which are employed to store previous search queries, previous search results, and downloaded resources and to allow users to access contributed searches. Also in development is the capability to display visualizations or previews for those resources that can provide them. Little work has been done to date on the user interface for nonscientific users, i.e., *SRAS administrators* and *resource providers*, who want to make their resources available through SRAS.

The Conceptual Model: Enabling the Developer

As noted earlier, the solar-terrestrial problem domain covers scientific exploration from the Sun to the Earth for over a decade of missions, observations, and experiments. A descriptive and intuitive model of this domain, including a significant level of abstraction, is critical to providing enough overall understanding to space scientists so that they can comfortably work in the domain and to computer scientists so that they can build systems to support the science missions. Several efforts are under way to address issues related to describing this domain, including the following.

ISO Standard 11179, *Information Technology: Specification and Standardization of Data Elements*,[15] provides guidelines for defining data elements and their metadata descriptions, along with a registration mechanism to support data interchange among people and information systems. The standard identifies metadata formats and data registration methods that providers of scientific data must follow in order to participate in a larger data access system. Despite the obvious benefit of an internationally accepted approach, this standard is particularly difficult to implement in the solar-terrestrial problem domain because so many sets of existing or legacy data were collected and developed under other programs or missions, many of which have ended and no longer have the funding necessary to generate metadata definitions for other programs. Furthermore, for both financial and scientific reasons, it would be challenging to require scientists participating in future LWS missions to generate metadata in a prescribed format. Finally, the standard focuses on detailed definitions of specific data items and does not support a robust mechanism for understanding the conceptual relationships among those data items. Thus, in systems that use ISO Standard 11179, data discovery and interchange occur at a very low level, requiring detailed user knowledge of the data concepts.

The Space Physics Archive Search and Extract (SPASE) system,[16] in which APL participates, is an international, multi-institutional effort to model the space physics domain and to build a system that enables data search and retrieval across participating data centers. This collaboration is tackling the difficult problem of integrating a variety of archive formats and access methods into a common interface, but focuses on a detailed modeling of a specific domain: space physics. The solar-terrestrial problem domain is broadly defined across scientific disciplines. It requires the capability to support queries at a number of very different levels of abstraction (e.g., "Find the particle flux for CNO nuclei at a particular sensor" or "Describe the measurable impact on Earth of a solar flare event on a specified day"). Nevertheless, the SPASE effort will certainly improve space physics data access and retrieval and contribute useful concepts to larger data integration systems such as the SRAS.

Modeling approaches for data integration systems tend to be characterized by the detailed description of the specific problem domain and a tight coupling to the available data resources. While this approach is successful for well-defined and easily scoped problem areas, it does not scale as both the data domain and the available resources increase in size and complexity, which is expected to be true in the LWS program. To address this concern, APL has developed a unified KB with three particular components, which are described next.

1. A high-level conceptual model of the scientific disciplines and domain data in the region from the Sun to the Earth
2. A detailed description of resource metadata supporting any accessible data resource
3. A loose coupling between the conceptual model and the resource metadata

1. High-level conceptual model. Rather than expressing the data environment in terms of each scientific concept and the relationships among them, which would yield an enormously large and complex model for the solar-terrestrial problem domain, a high-level conceptual model was developed from classes of scientific concepts and the general relationships among them (Fig. 6). Instances of these classes then represent specific domain concepts, and instances of the general relationships identify actual connections between particular scientific concepts. The model begins with the notion of *regions* that collectively define the area of interest for the LWS program, from the Sun to the Earth. Each region (enumerated as Solar, Interplanetary, Geomagnetic, and Near-Earth) contains *features* that exist for significant periods of time and *events* that occur at particular points in time. Within each region, events are related to features (e.g., the feature may define the location of an event, or an observation of the feature may be affected by the event). Furthermore, events can propagate through the regions, so an event in one region of interest could be related to an event in an adjacent region. Finally, all regions, features, and events are related to a set of *parameters* that define the working vocabulary for the Sun–Earth data domain.

2. Detailed description of resource metadata. Details about data resources must be readily available for a user to discover, access, and retrieve desired information. The resource metadata model in the SRAS KB includes descriptive information to help the user understand the data that are available and decide whether those data are appropriate for a particular inquiry. These *descriptive metadata* include program, satellite, and instrument information, as well as data details like observed phenomena, units of measure, time and spatial coverages, provenance, and processing level. In addition, *structural metadata* in the KB define the details for data access and retrieval if selected by the user. Historically, data resources have been maintained in file hierarchies, possibly accessible over a network by FTP or http. The notion of a data resource in APL's metadata definition also includes databases and computer programs (such as format conversion tools or data models) that generate requested data on demand. Thus, file access, database access, and program invocation details are all included in the resource metadata model.

3. Loose coupling. The conceptual model containing scientific understanding of the solar-terrestrial problem domain is integrated with the resource metadata describing available data and retrieval details in the SRAS KB by relating all data resources to the corresponding scientific parameters. In this way, a single resource can be related to all appropriate regions, features, and events simply by identifying the associated domain vocabulary.

APL's model design for the SRAS KB has a number of advantages over the single standard or detailed

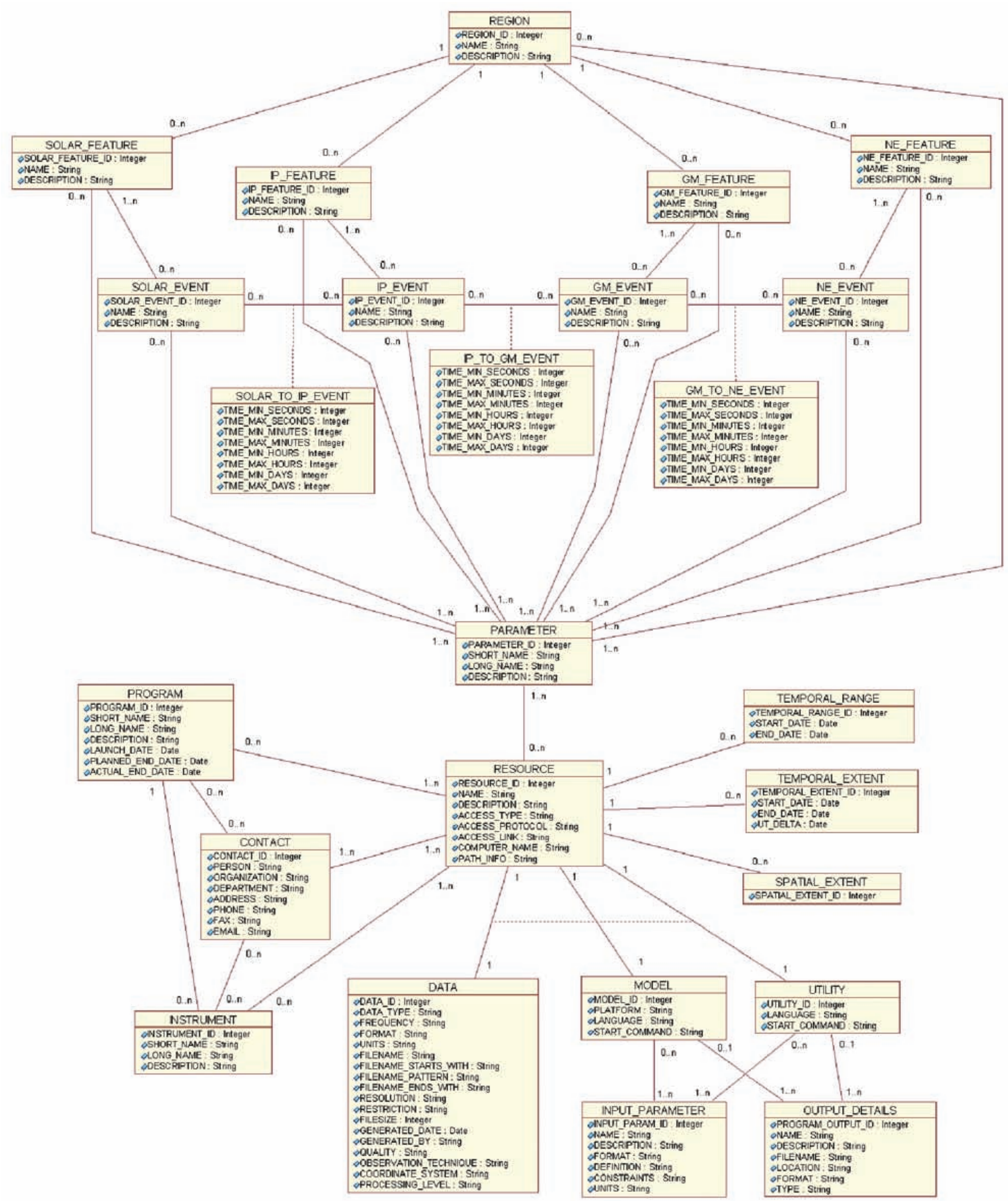

Figure 6. The SRAS conceptual model. An intuitive, abstract model of the problem domain enables advanced search features without being so complex that it becomes unwieldy.

modeling approaches previously mentioned for a general data discovery and retrieval system serving a variety of users. First, relating diverse data across scientific disciplines from disparate sources is easily accomplished through their scientific relationships. A query in the conceptual model identifies the desired parameters, which are then mapped to the related data resources. Second, data discovery is simplified by performing requests at the conceptual level, thereby permitting users without knowledge of specific mission, instrument, or observation details to readily find data of interest. Of course, expert user access to specific data from a particular source is also directly supported by the resource metadata details. Third, the conceptual model can be readily expanded if new classes of scientific concepts or new relationships among scientific concept classes would improve the model's representation of the Sun–Earth domain. Thanks to the high-level abstraction of the model, its implementation is unlikely to become too large and complex to support real-time query processing. Fourth, a data provider can add resources (of any format, access method, or processing level) simply by providing information that describes the data and the associated access details. The model eliminates the need for rigid metadata standards or unique, resource-specific integration software. Finally, the flexible mappings through the Parameters class ensures that any term that is useful in describing a scientific concept in the problem domain can be added to the query vocabulary. In this way, the model supports alternate terms derived from different scientific disciplines and facilitates successful searches for a variety of users.

The following examples illustrate the unique search and retrieval capabilities that are enabled through this approach:

- *Time-related data access*. Often a scientist searches various continuous sensor measurements—for example, geomagnetic indices such as Kp and Dst, solar wind ion parameters from the ACE Solar Wind Electron Proton Alpha Monitor (SWEPAM), and interplanetary magnetic field data from the ACE Magnetometer—to look for interesting activity. Normally, this task requires accessing the various data sources independently and comparing the data provided in separate application windows formatted on different presentation scales. The KB model supports a more convenient approach to this type of investigation in the form of a higher-level request for integrated time-series data. Individual data sets from different sources within the requested time interval are easily accessed and retrieved using the resource metadata details. Those data sets, in turn, provide the necessary input to a programmatic resource such as a plotting utility, which generates a combined graph of the requested data on a single time scale. Thus, with a single data request, the user receives a convenient, integrated representation of the desired data without manual intervention, permitting him or her to spend time and energy investigating scientific phenomena, not manipulating data.
- *Event-driven data discovery*. Events that occur on the Sun can have a measurable impact on the Earth, so a scientist might want to investigate the propagation of a solar event as it influences other regions of interest. In particular, consider a query to discover all data available for a coronal mass ejection (CME) that occurred at a specified time. The KB model provides direct relationships between events in adjacent regions, with additional attributes such as time ranges. For example, energetic particles from a CME will traverse the interplanetary region to be measured at L1 (the point in space where a spacecraft's orbital motion matches that of the Earth, enabling it to maintain its position relative to both the Earth and the Sun) within minutes, but it take days for the corresponding shock wave to arrive. With these relationships in the KB, a single high-level user request can discover a variety of available data—for example, CME images from the Solar and Heliospheric Observatory (SOHO), interplanetary particle flux data measured by ACE, near-Earth energetic particle plots from POLAR, and global convection maps from the Super Dual Auroral Radar Network (SuperDARN)—all selected within the appropriate time windows to produce related data for CME propagation.

While much has been accomplished in the development of this conceptual model, it is still a work in progress. Only a limited number of resources have been mapped into the model. As more are added, it is likely that additional concepts and possibly levels of abstraction will be required, in addition to the further development of new mapping techniques. Structural metadata will also evolve as additional resources are added to more fully integrate specific resource site capabilities. Finally, descriptive metadata will continue to be refined as scientists exercise the system and request information.

SUMMARY

The ultimate goal of any technical data access system is to reduce or even eliminate the barrier between the scientist and the desired scientific resources in order to make scientific data access, analysis, fusion, and synthesis as easy and intuitive as browsing the Web. Countless projects across scientific disciplines attempt to reach this goal by either tightly restricting the types of resources or by requiring rigorous adherence to a set of specifications. To our knowledge, the SRAS project at APL is unique in trying to reach this goal by loosely coupling highly heterogeneous resources to an integrated interface, minimizing the effort to integrate new resources, and maximizing the number of resources that can be made available though the interface.

Because the SRAS is fundamentally an integrator of scientific resources, it can take advantage of the success of other science resource projects, building on deployed data discovery and processing capabilities. Thus, resources accessible from the SRAS will include other data services that provide or simplify resource access, such as the Virtual Solar Observatory (VSO),[6] the Virtual Ionosphere Thermosphere and Mesosphere Observatory (VIO),[17] the Virtual Space Physics

Observatory (VSPO),[18] and the Space Physics Data Markup Language (SPDML) System.[19] Furthermore, the SRAS architecture framework permits integration in the opposite direction, so that broader science resource systems can integrate the SRAS as a single information discovery and access resource. In this way, all existing and new space science data projects can focus on their most critical contribution to the larger data problem (whether that is data discovery, data access, data processing, data integration, or something else), and every project can share in the cumulative benefits.

APL's progress to date includes a user interface that allows scientists to specify searches by physical parameters instead of resource-specific parameters, an innovative approach to organizing information about resources that enables both broad-based and specialized searches, and an extensible framework that provides hooks to integrate capabilities so that third parties can provide the visualization, data management, and data fusion tools needed by solar-terrestrial scientists to maximize the use of all available resources.

At least three major challenges lie ahead for the SRAS project.

1. To provide a clear, consistent, and useful physically based interface that can be used to specify searches across not just a dozen, but hundreds or thousands of resources
2. To greatly reduce the effort required to integrate new resources, particularly resources such as tools or models that dynamically generate data—a daunting algorithmic and user interface task, given the exceptionally wide variety of data formats, data access methods, and data models out there
3. To integrate tools and models into the request process itself, enabling the dynamic calculation of query parameters and corresponding resources to more accurately determine available scientific resources

Despite these serious challenges, however, it is strongly believed the SRAS concept of homogeneous access to heterogeneous resources through physically based searches and loosely coupled resources offers the right path toward the ultimate goal relatively simply and in a timely manner.

REFERENCES AND NOTES

[1]Withbroe, G. L., "Living With a Star," *Bull. Am. Astron. Soc.* **32**, 839 (2000), http://lws.gsfc.nasa.gov/lws_presentations.htm.

[2]Kozyra, J. U., Liemohn, M. W., Mlynczak, M. G., Paxton, L. J., Skinner, W. R., et al., "TIMED Observations of the Signatures of Magnetic Activity in the MLTI Region Placed into Global Context by ACE, POLAR, IMAGE, SAMPEX, FAST, NOAA/POES, and DMSP," Invited speech, *AGU Spring Mtg. 2002*, http://www.agu.org/meetings/sm02top.html.

[3]Semmel, R. D., Immer, E. A., Silberberg, D. P., and Winkler, R. P., "Knowledge-Based Query Formulation for Integrated Information Systems," *Johns Hopkins APL Tech. Dig.* **18**(2), 261–270 (1997), http://techdigest.jhuapl.edu/td1802/2semmel.pdf.

[4]TIMED mission Web site, http://www.timed.jhuapl.edu, and TIMED instruments Web sites: Global Ultraviolet Imager (GUVI), http://guvi.jhuapl.edu; Sounding of the Atmosphere using Broadband Emission Radiometry (SABER), http://saber.larc.nasa.gov/; Solar EUV Experiment (SEE), http://lasp.colorado.edu/see/; and TIMED Doppler Interferometer (TIDI), http://tidi/engin.umich.edu/.

[5]*TIMED Project Data Management Plan (DMP)*, 7363-9330, JHU/APL, Laurel, MD (2001), http://www.timed.jhuapl.edu/scripts/mdc_documents.pl.

[6]Davey, A. R., Bogart, R. S., Gurman, J. B., Hill, F., Hourcle, J., et al., "Implementation of the Virtual Solar Observatory," in *Proc. 204th Am. Astronom. Soc. Mtg.*, #70.02 (2004), http://umbra.nascom.nasa.gov/vso/.

[7]ModelWeb: Space Physics Models Web site, National Space Science Data Center (NSSDC), Goddard Space Flight Center, http://nssdc.gsfc.nasa.gov/space/model/.

[8]Bowshock Browser: Space Science Data Web site, NSSDC, Goddard Space Flight Center, http://nssdc.gsfc.nasa.gov/ftphelper/bowshock.html.

[9]Carr, S. S., Meng, C. I., and McMorrow, D. J., "University Partnering for Operational Support (UPOS) Space Environmental Projects," *AGU Spring Mtg. 2002*, http://www.agu.org/meetings/sm02top.html; UPOS Web site, http://sd-www.jhuapl.edu/UPOS/index.html.

[10]Satellite Situation Center Web (SSCWeb) service site, http://sscweb.gsfc.nasa.gov/.

[11]Coordinated Data Analysis Web (CDAWeb) service site, http://cdaweb.gsfc.nasa.gov/.

[12]Onsager, T. G., and Balch, C. C., "Operational Space Weather Products at the NOAA Space Environment Center," *AGU Spring Mtg. 2002*, http://www.agu.org/meetings/sm02top.html; the Space Environment Center Web site, http://www.sec.noaa.gov/.

[13]Coupling, Energetics and Dynamics of Atmospheric Regions (CEDAR) Web site, National Science Foundation, http://cedarweb.hao.ucar.edu/cgi-bin/ion-p?page=cedarweb.ion.

[14]Earth Observing System (EOS) Web site, Goddard Space Flight Center, http://eospso.gsfc.nasa.gov/.

[15]*Information Technology: Specification and Standardization of Data Elements, Parts 1–6, ISO/IEC 11179*, International Organization for Standardization/International Electrotechnical Commission, Joint Technical Committee 1—Information Technology Standards, Sub-Committee 32—Data Management and Interchange, Working Group 2—Metadata (ISO/IEC JTC1 SC32 WG2) (1995–2004).

[16]Thieman, J. R., King, T., Roberts, A., and Walker, R., "SPASE—The Space Physics Archive Search and Extract—Status and Plans," *AGU Fall Mtg. 2003*, http://www.agu.org/meetings/fm03/; SPASE Web site, http://www.igpp.ucla.edu/spase/index.cfm.

[17]Yee, J.-H., Talaat, E. R., and Nylund, S. R., *A White Paper Plan for the Definition of a Virtual ITM Observatory (VIO)*, JHU/APL, Laurel, MD (2003).

[18]Rezapkin, V., Roberts, D. A., Coleman, J., and Boller, R., "Building a Virtual Space Physics Observatory for Easy Access to and Novel Visualization of Distributed Data," *AGU Fall Mtg. 2003*, http://www.agu.org/meetings/fm03/; Virtual Space Physics Observatory (VSPO) Web site, http://vspo.gsfc.nasa.gov/websearch/html/VSPO.html.

[19]Weiss, M. B., Morrison, D. D., Hashemian, M. R., Barnes, R. J., and Paxton, L. J., "Space Physics Metadata Searching Using Space Physics Data Markup Language (SPDML)," *AGU Fall Mtg. 2003*, http://www.agu.org/meetings/fm03/; SPDML Web site, http://sd-www.jhuapl.edu/SPDML/.

THE AUTHORS

Rose A. Daley is the Technical Lead for SRAS. Ms. Daley is the Supervisor of the Information Systems Engineering Section in APL's Research and Technology Development Center (RTDC) and focuses primarily on engineering and architecture challenges of information systems. **Elisabeth A. Immer**, a Senior Computer Scientist in RTDC, leads the knowledge-base design and data modeling efforts for SRAS. **Brand I. Fortner**, a Senior Professional Staff member in APL's Power Projection Systems Department (PPSD), works on user interface design issues for the SRAS team. Dr. Fortner is currently Chief Scientist of the Information Exploitation Group (VIE), where he works on imaging, video, and scientific data visualization projects. **Michele B. Weiss**, a Senior Professional Staff member in APL's Space Department, is the Software Lead for the SRAS team, and has more than 8 years of experience addressing challenges in providing relevant scientific information to space scientists. Further information on SRAS at APL can be obtained from the Technical Lead. Her e-mail address is rose.daley@jhuapl.edu.

Rose A. Daley

Elisabeth A. Immer

Brand I. Fortner

Michele B. Weiss