SPIDR middleware for WDCs

M. N. Zhizhin[1] and E. A. Kihn[2]

Received 17 September 2007; revised 18 March 2008; accepted 19 March 2008; published 21 March 2008.

[1] SPIDR (Space Physics Interactive Data Resource) is a de facto standard data source for solar-terrestrial physics, functioning within the framework of the ICSU World Data Centers. It is a distributed database and application server network, built to select, visualize and model historical space weather data distributed across the Internet. SPIDR can work as a fully-functional web-application (portal) or as a grid of web-services, providing functions for other applications to access its data holdings. *INDEX TERMS:* 0525 Computational Geophysics: Data management; 0530 Computational Geophysics: Data presentation and visualization; 7894 Space Plasma Physics: Instruments and techniques; 7959 Space Weather: Models; 7999 Space Weather: General or miscellaneous; *KEYWORDS:* Data Grid, phenomena-based subsetting, space weather, satellite data archive.

Citation: Zhizhin, M. N. and E. A. Kihn (2008), SPIDR middleware for WDCs, *Russ. J. Earth. Sci., 10,* ES3001, doi:10.2205/2007ES000281.

Introduction

[2] The Space Physics Interactive Data Resource (SPIDR) (http://spidr.ngdc.noaa.gov) is a standard data source for solar-terrestrial physics, functioning within the framework of the World Data Centers. It is a distributed database and application server network, built to select, visualize and model historical space weather data distributed across the Internet. SPIDR can work as a fully-functional web-application (portal) or as a Grid of web-services, providing functions for other applications to access its data holdings.

[3] Currently SPIDR archives include solar activity and solar wind data, geomagnetic, ionospheric, cosmic rays, radiotelescope ground observations, telemetry and images from NOAA, NASA, and DMSP satellites. SPIDR portals, databases and services are installed in the USA, Russia, China, Japan, Australia, South Africa, India, France and Ukraine. SPIDR has more than 20 000 registered world-wide users and daily load of about 100 user sessions per site. SPIDR technology has proven to be useful for environmental data sharing, visualization and mining, not only in space physics, but also in diverse environmental arenas such as seismology, GPS measurements, tsunami warning systems, and others.

Background and Related Work

[4] SPIDR is a type of Grid for environmental data. We define a Grid of environmental data sources to be a set of web services following the same contract for dynamic service registry, metadata and data request interfaces, as well as output metadata scheme and data model. This is inline with the general Grid approach towards virtualization of data, services, and interfaces [*Zhao et al.,* 2006]. "Behind" the web service we can store the environmental data in a file system as binary files or images, in a relational database as rows of observations, or as another web service possibly with a different service contract. Each storage method and structural organization of a dataset will require a specific implementation of our "virtual" data source web service, but for the user of the Grid all of them will look like the same Common Data Model (CDM) apart from the specific environmental data contents, such as parameters, stations, grid–coordinates and observation time intervals.

[5] We have been developing this concept of a virtual environmental data source for some time already, starting with distributed web services and portals for the space physics, meteorological and simulation communities. There are two main reasons why are moving to Grid middleware and infrastructure, which is more complicated for development and support as compared to a "pure" web-services approach implemented in the "standard" Apache Axis or Microsoft .NET web-services container:

- The availability of a scalable Virtual Organization proxy mechanism for individual users based on digital certificates used by Grid for secure access to multiple distributed resources compared to the local portal user-password authentication [*Foster et al.,* 2001];
- A data request and/or processing from large environmental archives may take quite a while even if we specifically optimize the database structure and the processing algorithms for this type of request, and a

[1]Geophysical Center RAS, Moscow, Russia
[2]National Geophysical Data Center, Boulder, Co., USA

ISSN: 1681–1208 (online)

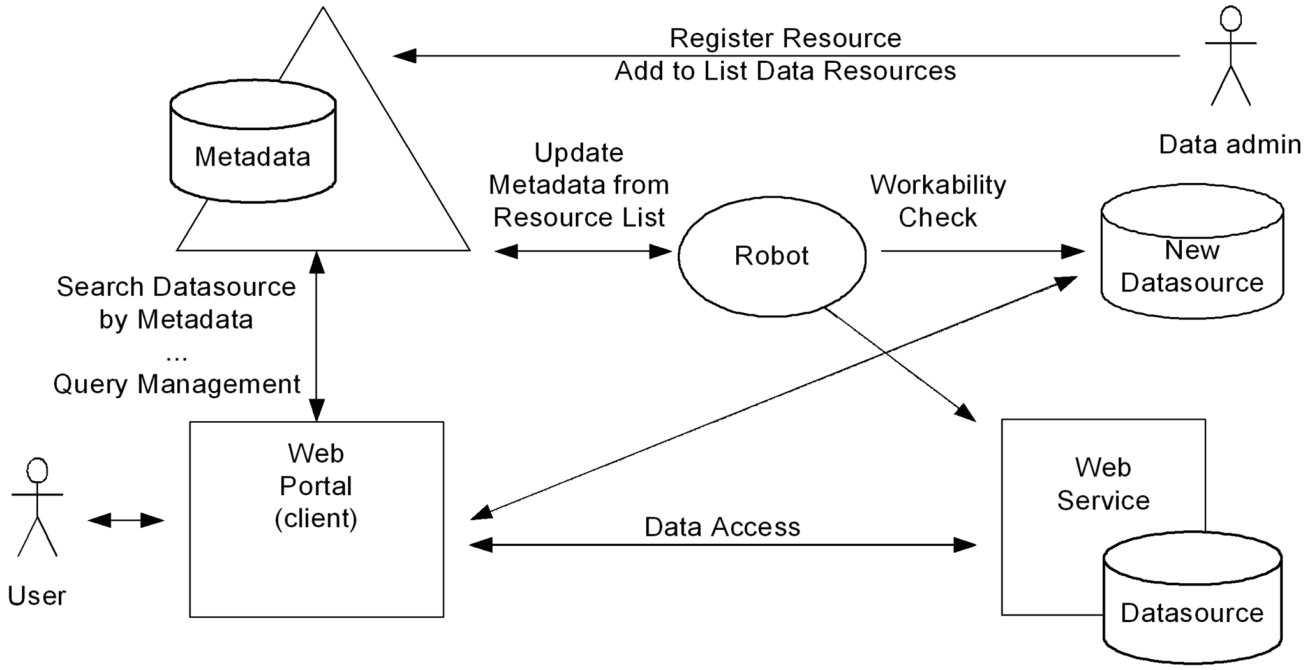

Figure 1. A web-portal as a client for a Grid of data sources.

synchronous web-services call mechanism is not always appropriate to handle data requests which involve a long time delay [*Barkstrom et al.,* 2003].

[6] The SPIDR system concept is similar to several emerging technologies for data access in the environmental sciences. Notable among these are Unidata's Thematic Real-time Environmental Distributed Data Service (THREDDS) [*Domenico et al.,* 2006], the Environmental Scenario Generator (ESG) from the USAF [*Kihn et al.,* 2004], and the Coordinated Data Analysis Web (CDAWeb) from NASA (http://cdaweb.gsfc.nasa.gov).

System Architecture

[7] The SPIDR system architecture has the following main components: a web-portal, metadata repository, visualization and data mining engines, and a grid of virtual data sources exposed to the external clients including the SPIDR portal via data query and inject web services. Behind a data source's web service one can have a database, a set of files in a local to server file system, or a set of URLs to remote data sources.

SPIDR Portals

[8] A web-portal serves as an agent between the user and the Grid of environmental data sources. It performs two main functions. The first function is metadata management, which allows for fast and efficient catalog-level metadata search. Here by catalog-level metadata we mean general descriptions of data resources, stored as a managed collection of XML documents with a known XML schemas (i.e.. owner info, geographic coverage, time coverage, data description, visualization methods, etc.). Our catalog-level metadata collection works much the same as other similar resources, e.g. Global Change Master Directory (GCMD) from NASA (http://gcmd.nasa.gov).

[9] The second function of the web-portal is data access. In Figure 1 the web-portal is shown as a client, which connects to virtual data sources, retrieves the requested data, and delivers it back to the user. Advanced web-portal functions can include visualization and data mining. Data access web forms are built using inventory (or granule-level) metadata describing the availability of stations – satellites – instruments – parameters or channels for the given time interval. The inventory metadata can be used also to compare and synchronize mirrored data sources.

[10] The SPIDR portal combines a central metadata registry with a set of distributed data web services, web map services, and replica sets of data files. A user can search catalog-level metadata and inventory, use a persistent data basket to save the selection between the sessions, and plot or download the selected data in different formats, including XML and netCDF. A database administrator can upload files into the SPIDR databases using either a web services or web portal interface.

Metadata Registry and Data Inventory

[11] Both the catalog- and granule-level metadata records, which contain respectively a general description and detailed

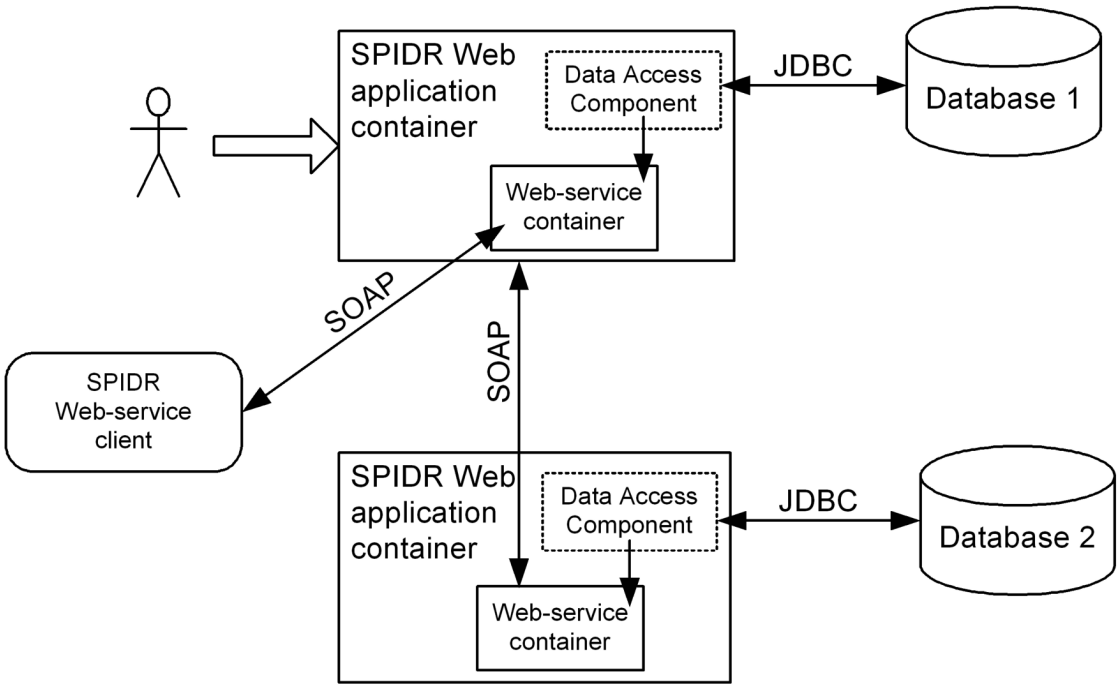

Figure 2. Local vs. remote SPIDR data sources.

inventory of SPIDR data resources, can be updated either manually by a system administrator or automatically by the data robot collecting records from the data grid (see Figure 1). The catalog-level metadata registry uses a native XML database backend based on the open-source product eXist [*Meier,* 2006]. The metadata engine has no predefined XML schema; it is possible to have different metadata schemas for different data categories. For example, data sources with spatial content, such as OpenGIS Web Map Services (http://www.opengeospatial.org/standards/wms) and time series databases with ground observations, can use FGDC Content Standard for Digital Geospatial Metadata (document FGDC-STD-001-1998, http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/index_html), and at the same time databases with satellite telemetry can have SPASE-formatted metadata records (Space Physics Archive Search and Extract, 2006, http://www.spase-group.org/data/). The SPIDR high-level metadata engine has extended search capabilities allowing it to search in specific metadata elements, (keyword, title, provider, etc.). In addition, it supports Web 2.0 style functionality with direct editing of the metadata records at the SPIDR portal, a user discussion forum, internal e-mail messaging, and wiki-style documentation and system help.

[12] The SPIDR granule-level inventory metadata registry uses an SQL database backend based built on the open-source product MySQL (http://dev.mysql.com/doc/refman/5.0/en/index.html, 2008). The main purpose of the inventory is to list available parameters and stations from each database with some granularity in time, currently taken as monthly. That is whether a given station has any data for a given month. This information is needed in early validation of data requests for both availability and size of the data export, and for comparison of data holdings at different SPIDR nodes for database synchronization. When adding new data to SPIDR, the inventory can be updated either in real time or by periodic queries of the corresponding data source, depending on the input data load. At the same time the inventory metadata is updated the inventory summary such as the station and parameter list with maximum date ranges is fed up in order to update the corresponding catalog-level metadata.

Grid of Web-services

[13] Web Services (WS) technology is used by SPIDR to access databases and metadata both for the SPIDR web application (interactive interface for human users) and for the SPIDR web clients (third party programs exporting and importing data and metadata in batch mode). In addition to the WS SOAP protocol the SPIDR web application can access databases directly using JDBC drivers. We call the JDBC-connected databases "local" and the WS-connected databases "remote" (Figure 2). The access mode is defined in the database configuration files. If database is hosted on the server on the same LAN as the SPIDR web application, then the local access mode may be more efficient compared to the remote one; but if the database is located outside the local network then the JDBC connections will be the most probably blocked by a security considerations and the SOAP protocol becomes the only reliable way to access the data.

[14] SPIDR data archives are logically organized into thematic groups called viewGroups (e.g. Geomagnetic Indices,

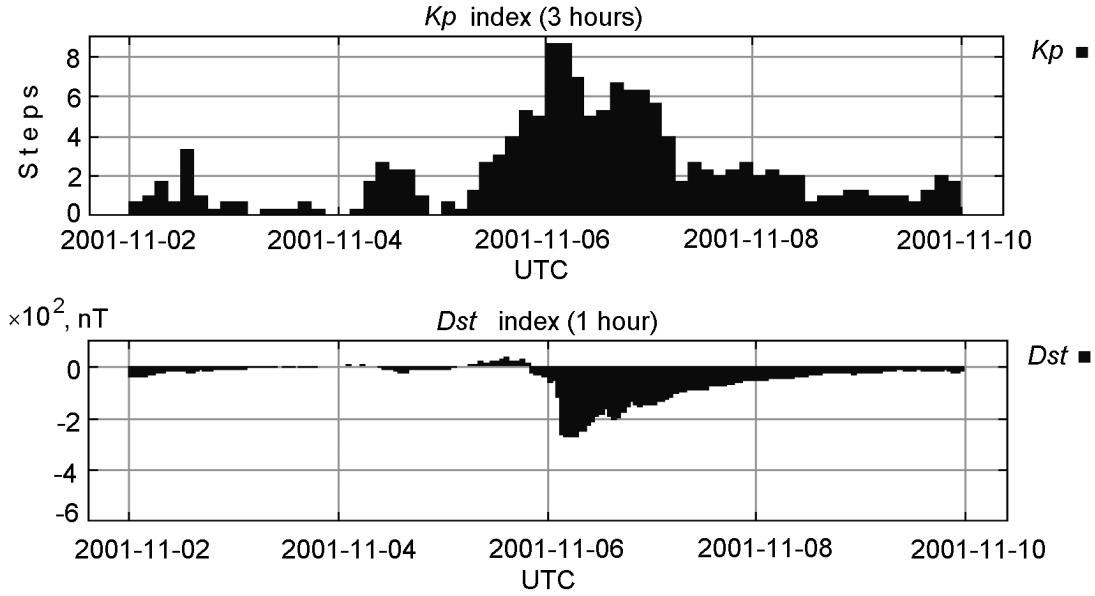

Figure 3. Simultaneous plots of data selected from different services.

or Interplanetary Magnetic Field). Each viewGroup may include several databases or groups of database tables, which we call tables. Each table may be considered as a virtual database with a single configuration file describing the access mode (local or remote) together with URL and access credentials.

[15] The system currently has implemented web-services for the following use cases:

1. Get a metadata record for a given viewGroup (Virtual Observatory WS).

2. For a given table, element, station and date interval get a data inventory and export data values in a variety of scientific formats, including XML and NetCDF (Data Source WS).

3. Load several "standard" data files of several scientific formats into the database (Data Sink WS).

4. Synchronize two SPIDR archives by exporting data from one archive and loading into another (WS orchestration).

[16] **Data Source Web-service.** Data source web service URLs are stored in SPIDR configuration files. When a web service call is performed, the web-service returns a URL, pointing to a data file, containing the serialized CDM object with requested data. The SPIDR application itself can act as a web-service and process remote calls thus allowing for chaining of the data export web services.

[17] In any case, a SPIDR data source service will supply metadata describing parameter names, units of measure, visualization options, etc., and the data accreditation describing the data origin. Data serialization formats include direct Java object serialization, XML, NetCDF, and for some databases also special formats introduced by the data users community. For example, geomagnetic field variations can be exported in WDC or Intermagnet formats. For geomagnetic variations the data accreditation describes the observatory which has provided the data to SPIDR.

[18] With the SPIDR portal, a user can collect the serialized data from the distributed data sources into a single "user basket", re-format and re-package all the data for download, or visualize the selection with multiple time-synchronous plots either using static GIF images or by dynamic "zoomable" Java applets which share the same time scale limits. In Figure 3 we present an example plot of selections from two databases with planetary geomagnetic disturbance index Kp and Disturbance Storm Time index DST. Both indices are prime indicators of a magnetic storm, and the simultaneous plots help to estimate the storm intensity.

[19] Data query options (time interval, data source, parameters, stations) are saved in the user basket, so the data can be re-selected in the future. Because of the real-time nature of the SPIDR databases, the data selection itself is transient, so theoretically in the next session user can find different (updated) observations in the data basket. All the data selection queries are logged, so the SPIDR administrator can view not only user session statistics, but also the frequency of data requests by source.

[20] **Satellite Data Granules and Image Archive Web-services.** Remote sensing and imagery databases have a different data model as compared to a sequential database. Usually the data collection is divided into "elementary" blocks called granules. A granule can be a daily set of solar images from different observatories, or a fixed-length section of satellite orbit with Earth observations in different spectral bands.

[21] For example, the magnetic storm shown by the time series plots in Figure 4 is manifested in the daily solar image granule by a bright solar flare in the 164 MHz radiotelescope

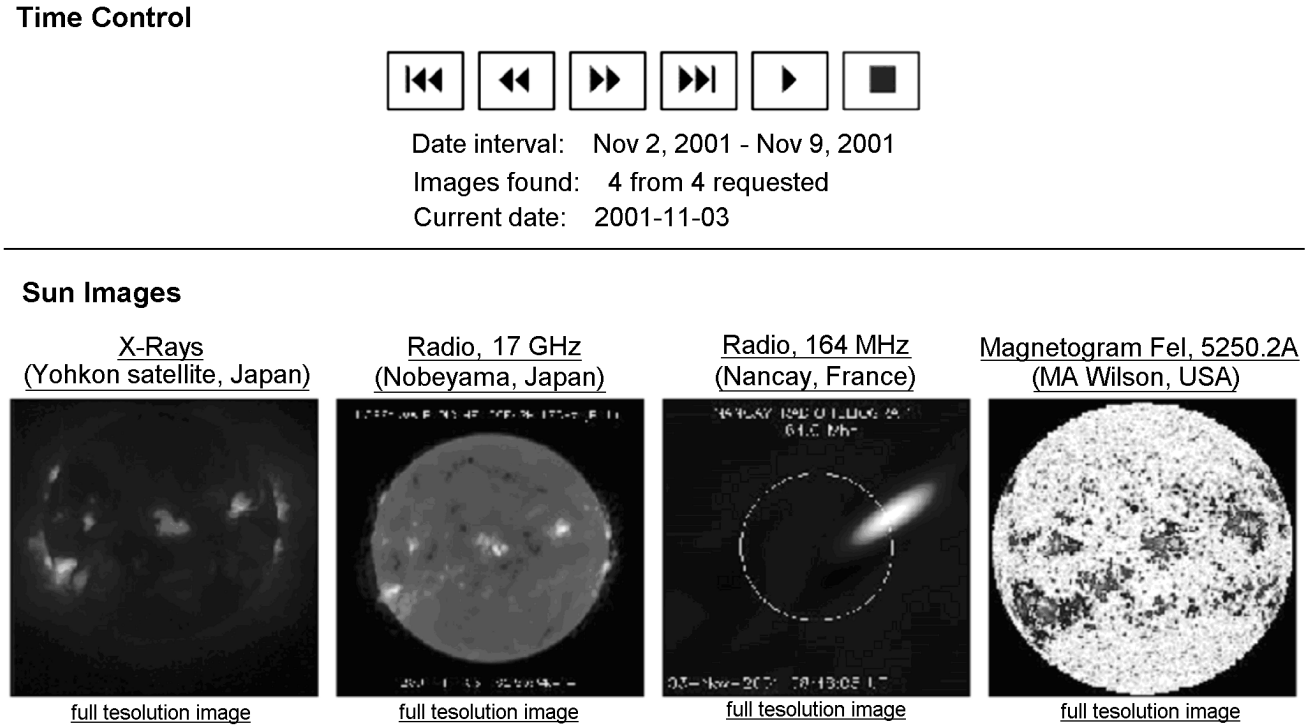

Figure 4. SPIDR browser for solar image granules with a visible geo-efficient flare.

image. The flare erupts from a large system of sunspots visible on the solar X-ray and magnetogram images. At the same time, the aurora produced by this magnetic storm on the night side of the Earth can be seen at the cloud-free night-time image granules of 1/8th of the DMSP satellite orbit; one of them is shown in Figure 5.

[22] All the granule-based web-services in SPIDR have the same design pattern. The user's data export request specifies the date range and type of the image. The web-service returns a list of granules with metadata and links to the preview and high-resolution images or binary files for granule data products like DMSP satellite SSJ/4 sensor readings.

[23] **Data Sink Web-service.** Clients can load data into SPIDR databases from files located on a local workstation, along with relevant loading options which are passed to the SPIDR web services together with the data over SOAP with attachments. The database loading web service called by the client will parse the input file format, load data into the local database, add a bookkeeping record into the SPIDR data input/output logging database, and check the list of mirrored SPIDR nodes to send the input data file there to keep those databases in sync.

Database Synchronization

[24] SPIDR databases are self-synchronizing (Figure 6). The synchronization has both push and pull modes and it is based on the data source and the data sink web-services. In the push mode, when a new data set is successfully parsed and loaded into a database at one of the SPIDR nodes (we call it "master") using the data sink web service, all other nodes which are subscribed to this data stream (we call them "slaves") will receive the same set of data exported from the "master" node using the data source web service. Each SPIDR node can be either "master" or "slave" depending on whether it receives data from external sources or from another node. Such a peer-to-peer synchronization via web-services CDM object exchange has many advantages for heterogeneous distributed system, where SPIDR nodes can run different operating systems, database engines, and network security policies. For a high volume of short input messages, we can use pull mode synchronization. In this case the "slave" node periodically calls the "master" data source and receives, say, the last day of observations as a single data set.

[25] The SPIDR admin web interface has special tools to compare the same databases from several nodes and if necessary to order background synchronization from/to any of them. The inventory-level metadata from the "master" and "slave" nodes can be used to compare the data holdings and when there are any differences to start a background process at the "slave" node, which will pull the locally missing data from the "master" node using its data source web-service and load it into SPIDR by using the local data sink web-service.

[26] This web-services based synchronization mechanism is a new step in automation of the data exchange between

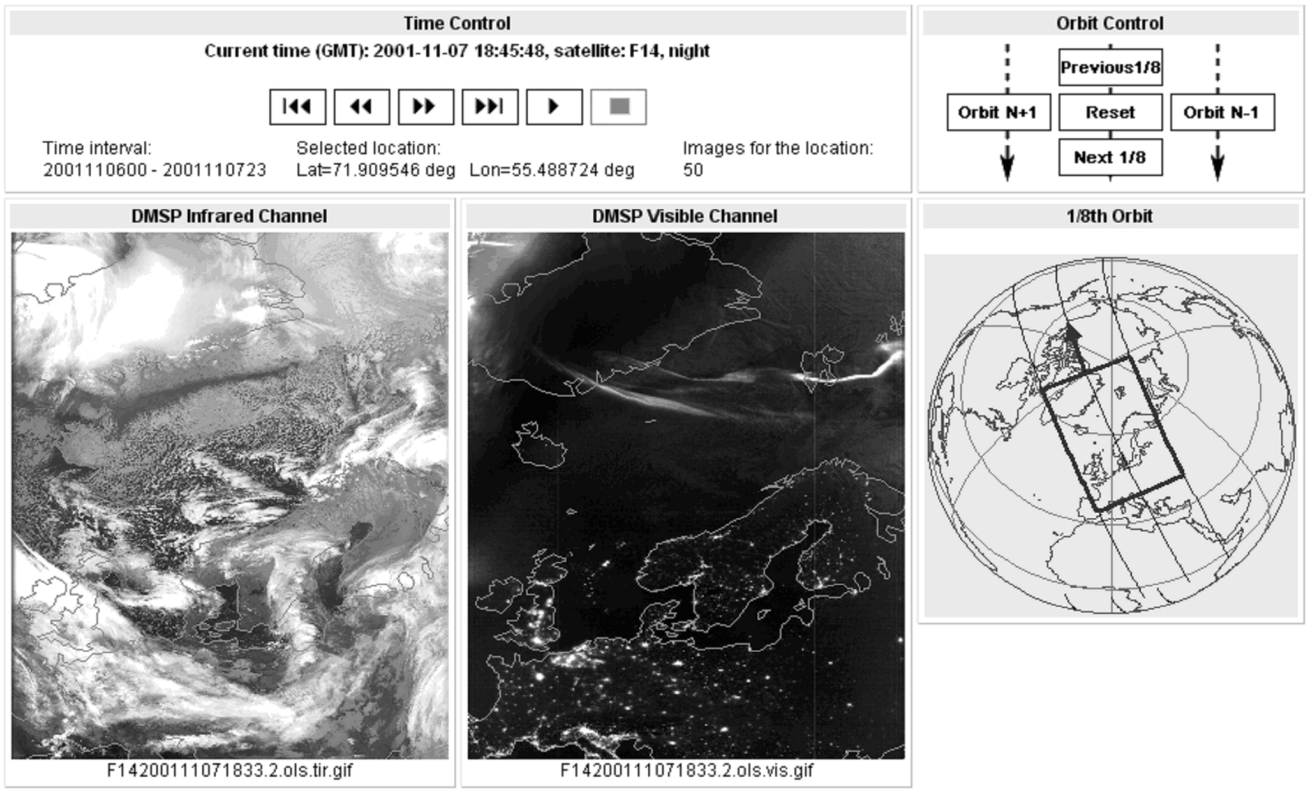

Figure 5. SPIDR browser for DMSP image granules.

World Data Centers in different countries. The common data model used by all the SPIDR nodes eliminates unnecessary format translations when synchronizing databases at different nodes. The peer-to-peer push synchronization aligns with agency priorities by first loading data into the national "master" node and then exporting the data to a given list of subscribers abroad. Existence of several copies of the same database in a very distributed network helps ensure long term data preservation.

Virtual Observatory

[27] A virtual observatory (VxO) a term now appearing within the scientific data community is a distributed software system that allows users to find, access, and use resources from a collection of data repositories and service providers. A virtual observatory can provide either metadata or data services and is typically focused on presenting the collection of data, metadata and functional services to a given set of customers bound by a common interest. The virtual observatory is an implementation of what is typically called service oriented architecture (SOA) bound by a common theme.

[28] Within the environmental community it is envisioned that using the VxO domain scientists will be able to:

- execute advanced search environmental archive queries based not only on metadata but on the included data content;
- conduct content-based query and data retrieval from virtual observatories.
- generate on-the-fly products interactively using existing data and metadata, as well as conducting detailed analysis;
- expand their ability to use and incorporate data from disciplines other than their own.

[29] As shown in Figure 7 the virtual observatory consists of a number of services built around a community based core.

[30] Because SPIDR is built around the Grid core it is easily adaptable to support the VxO infrastructure. SPIDR itself has been chosen as the basis of or a component of a number of virtual observatories. An example is the Virtual Radiation Belt Observatory (ViRBO) which is focused on the physics and phenomena surrounding the high energy particle belts surrounding the Earth and provides data from satellites and models covering the region.

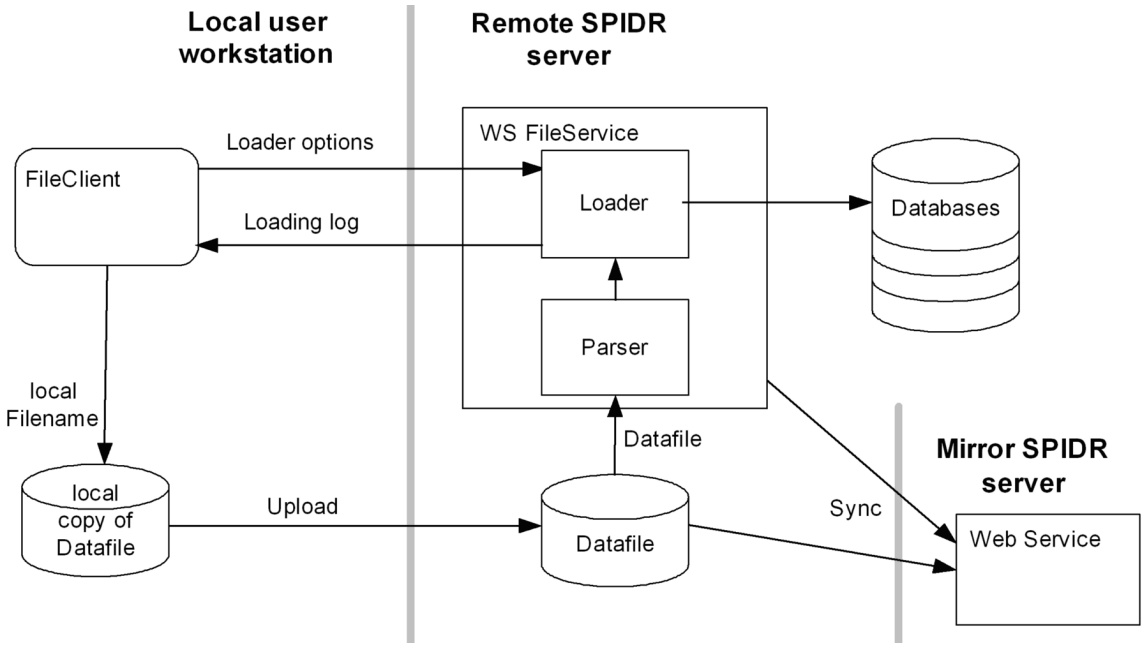

Figure 6. Data loading and synchronization SPIDR web service.

[31] In order to meet the "spirit" of a VxO, a system must encompass certain functionality. For example in the workflow component a system must:

- Support many possible levels of user interaction
- Support community-centric views including
  - community newsfeeds
  - portals to other VxO's
  - community specific tutorials
  - forum and wiki capabilities
  - software libraries and downloadable tools.

[32] The SPIDR interface meets all of these criteria by using the Jakarta Struts framework to define workflow. This

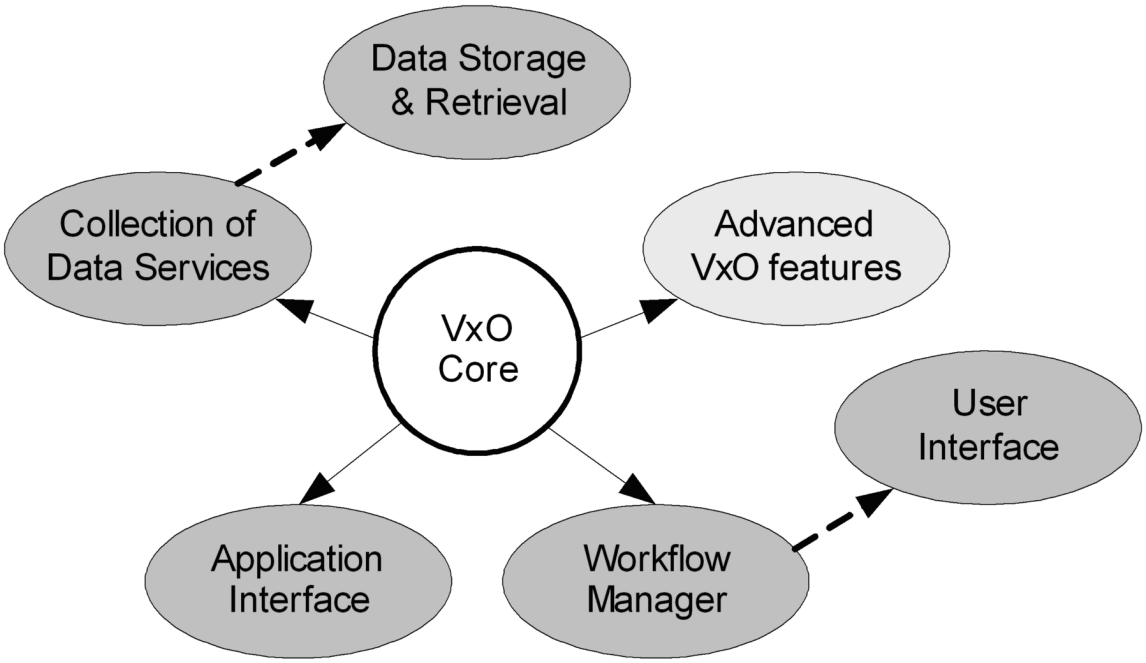

Figure 7. Typical Virtual Observatory services.

makes it possible to easily create multiple systems views focused on a particular user skill level (e.g. novice, advanced, admin) or discipline (e.g. geomagnetic, iononspheric, cosmic rays) without having to rewrite code. The interface can largely be adapted by editing XML documents from the Struts workflow configuration.

[33] Another element to a VxO is the Application Interface, here the system must provide:

- web-services based data flow and data transformations;
- API to interface with other VxO's;
- Common Computational and Data Model;
- mechanisms for creating derived data products;
- events of interest, e.g. magnetic storm detection.

[34] Obviously many of these items map directly to the Grid paradigm basically making any VxO an implementation of Grid. It is likely that as both progress there will be a merging in several key areas and particularly in environmental science related activities.

Conclusions

[35] It is our belief that increasing data volumes demand new tools and methods to maximize scientific efficiency and that software tools and mathematical methods exist which, provide analysis, classification, access, discovery and forecast methods for large volume data sets. Grid will play an important part in making these tools available on the internet for use with the distributed archives that are being developed now and in the future. The SPIDR system is an early implementation of a Grid system, which while discipline focused, exhibits the key operational components of a of a true Grid environmental data system. The SPIDR system itself may be used as a pattern for those interested in implementing such an environmental tool.

References

Barkstrom, B. R., T. H. Hinke, S. Gavali, W. Smith, W. J. Seufzer, C. Hu, and D. E. Cordner (2003), Distributed Generation of NASA Earth Science Data Products, *Journal of Grid Computing, 1,* 101, online at http://www.nas.nasa.gov/News/Techreports/2005/PDF/nas-05-006.pdf.

Domenico, B., J. Caron, E. Davis, R. Kambic, and S. Nativi (2002), Thematic Real-time Environmental Distributed Data Services (THREDDS): Incorporating Interactive Analysis Tools into NSDL, *Journal of Digital Information, 2*(4), 114, online at http://jodi.tamu.edu/Articles/v02/i04/Domenico/.

Foster, I., C. Kesselman, and S. Tuecke (2001), The Anatomy of the Grid: Enabling Scalable Virtual Organizations, *International Journal of High Performance Computing Applications, 15*(3), 200, doi:10.1177/109434200101500302.

Kihn, E. A., M. Zhizhin, R. Siquig, and R. Redmon (2004), The Environmental Scenario Generator (ESG): a distributed environmental data archive analysis tool, *Data Science Journal, 3,* 10, doi:10.2481/dsj.3.10.

Meier, W. (2006), Index-driven XQuery processing in the eXist XML database, XML Prague Conference Proceedings, Sourceforge, http://exist.sourceforge.net/xmlprague06.html.

Zhao, Y., M. Wilde, I. Foster, J. Voeckler, J. Dobson, E. Gilbert, T. Jordan, and E. Quigg (2006), Virtual data Grid middleware services for data-intensive science, *Concurrency and Computation: Practice & Experience, 18*(6), 595, doi:10.1002/cpe.968.

E. A. Kihn, National Geophysical Data Center, Boulder, USA (Eric.A.Kihn@noaa.gov)

M. N. Zhizhin, Geophysical Center RAS, Moscow, Russia (jjn@wdcb.ru)