Earth Sci Inform (2008) 1:79–91
DOI 10.1007/s12145-008-0012-5

METHODOLOGY ARTICLE

Space Physics Interactive Data Resource—SPIDR

Mikhail Zhizhin · Eric Kihn · Rob Redmon · Dmitry Medvedev · Dmitry Mishin

Received: 29 October 2007 / Accepted: 20 June 2008 / Published online: 16 July 2008

Abstract SPIDR (Space Physics Interactive Data Resource) is a 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.

Keywords Grid data mining · Phenomena-based subsetting · Product generation · Satellite data · Space weather

Introduction

The World Data Center (WDC) system (ICSU 1996) was created by the International Council of Scientific Unions (ICSU) to archive and distribute data collected from the observational programs of the 1957–1958 international geophysical year (IGY). Originally established during the IGY in the United States, Europe, Russia, and Japan, the WDC system has since expanded to other countries and to new scientific disciplines. The WDC system now includes 52 centers in 12 countries. Its holdings include a wide range of solar, geophysical, environmental, and human dimensions data. These data cover timescales ranging from seconds to millennia and they provide baseline information for research in many ICSU disciplines, especially for monitoring changes in the geosphere and biosphere—gradual or sudden, foreseen or unexpected, natural or man-made.

Since the IGY, technological advances have transformed the gathering and exchange of data. Starting with paper tables and magnetic tapes, and total data holdings of about ~1 Gb and annual data traffic ~1 Mb/year, in the early-nineties the WDCs have switched to Internet ftp and http protocols for regular environmental data exchange on a global scale with digital archives of ~1 Tb size and traffic ~1 Gb/year. We project that for the electronic geophysical year (eGY) in 2007 the world-wide Grid of WDC digital archives will reach petabyte size with terabyte-scale annual network traffic.

The Space Physics Interactive Data Resource (SPIDR) (http://spidr.ngdc.noaa.gov) originally developed in 1995 as a demonstration for the international Global Observation and Information Network (GOIN) project is now a standard data source for solar-terrestrial physics, functioning within the framework of the World Data Centers for Solar-Terrestrial Physics. SPIDR has gone through a total of four versions including a complete rebuild using exclusively open-source tools between version one and two. It is a distributed database and application server network, built to select, visualize and model historical space weather data spanning hundreds of years and 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.

Currently SPIDR archives include solar activity and solar wind data, geomagnetic, ionospheric, cosmic rays, radio-telescope ground observations, telemetry and images

Communicated by: Hassan Babaie

M. Zhizhin (✉) · D. Medvedev · D. Mishin
Geophysical Center, Russian Academy of Sciences,
Moscow, Russia
e-mail: jjn@wdcb.ru

E. Kihn · R. Redmon
National Geophysical Data Center NOAA,
Boulder, CO, USA

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 customers are predominantly academic and U. S. government users but a reasonable 20% also come from the commercial sector. SPIDR data and technology has application in 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

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 in line 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 (Fig. 1). 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. CDM for SPIDR was developed in parallel with the CDM efforts for CDF by NASA and NetCDF by UNIDATA; although different, it can be mapped into both of the "sister" data models.

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 the system is moving to Grid middleware and infrastructure, which requires a greater investment of "spin-up" time for development 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 and authentication with multiple distributed resources compared to the local portal user-password authentication (Foster 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 synchronous web-services call mechanism is not always appropriate to handle data requests which involve a long time delay (Barkstrom et al. 2003).

The SPIDR system concept is similar to several emerging technologies for data access in the environmental sciences. Notable among these are Unidata's Thematic

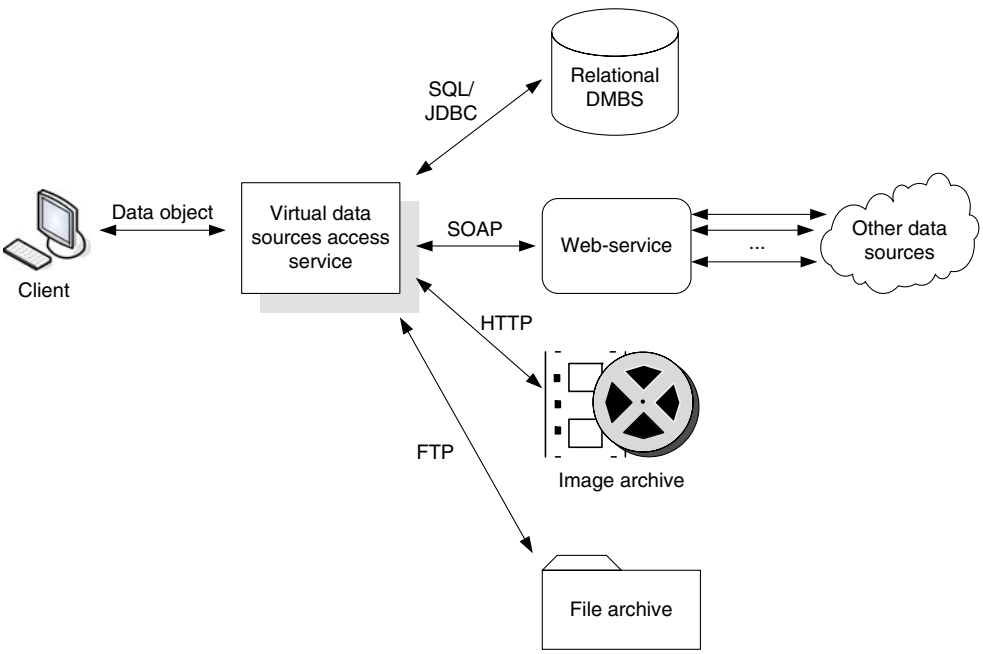

Fig. 1 Web services as mediators to the grid of environmental data archives

Real-time Environmental Distributed Data Service (THREDDS) (Domenico et al. 2002), 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).

THREDDS, like SPIDR, allows for data held in remote repositories, but it supports a different set of formats in that storage (notably netCDF). Both THREDDS and CDAWeb are file-system based repositories; SPIDR supports integration via plugin-based data virtualization from databases and data file repositories. Once an archive is made available in SPIDR, THREDDS or the CDAWeb system, it is automatically cataloged to generate metadata and that metadata is made available in an XML format to support cross system searching.

These data access systems use a common middle layer computational data model. That is to say the data is mapped from its storage format to a standard model which is used by the functionality in the middle layer (visualization, data mining, sub-setting, etc.). The THREDDS server uses the OpenDAP data model (Gallagher et al. 2007), while the ESG uses the Five Dimensional Representation (FDR) and the CDAWeb uses data model imposed by the Common Data Format (CDF) (http://cdf.gsfc.nasa.gov). All these data models can be mapped to each other (with some reservations) and support a similar pattern to the SPIDR data, that is representing the data in a time, space, parameter based model. The web services architecture style used by the data access layer can be either REST[1] (THREDDS) (Fielding 2000) or SOAP (ESG, CDAWeb) (Loughran and Smith 2005).

System architecture

The SPIDR system architecture has the following main components: a web application (portal), metadata registry (Virtual Observatory), visualization and data mining engines, and a Grid of virtual data sources, which are 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 server file system, or a set of URLs to remote data sources. SPIDR is an Open Source project (http://sourceforge.net/projects/spidr).

SPIDR portals

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 including owner info, geographic coverage, time coverage, data description, visualization methods (FGDC 1998). Our catalog-level metadata collection works much the same as other similar resources, including Global Change Master Directory (GCMD 2008) from NASA (http://gcmd.nasa.gov) or Master Environmental Library (MEL) from the US Defense Modeling and Simulation Office (Siquig and Lowe 1996). For a more detailed discussion on the role of metadata in distributed data networks see (Nieto-Santisteban et al. 2004).

The second function of the web-portal is data access. In Fig. 2 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.

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 ASCII text files, XML[2] 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

Both the catalog- and granule-level metadata records, which contain respectively a general description and detailed 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 Fig. 2). 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

[1] Representational state transfer (REST) REST is a web-service architecture style that exploits the existing technology and protocols of the Web, including HTTP and XML. REST is simpler to use than the SOAP (Simple Object Access Protocol) approach, which requires writing or using a provided server program (to serve data) and a client program (to request data). SOAP, however, offers potentially more capability.

[2] SPIDR simple XML data export schema has a semantic header followed by time-value element pairs similar to a pen-plotter language.

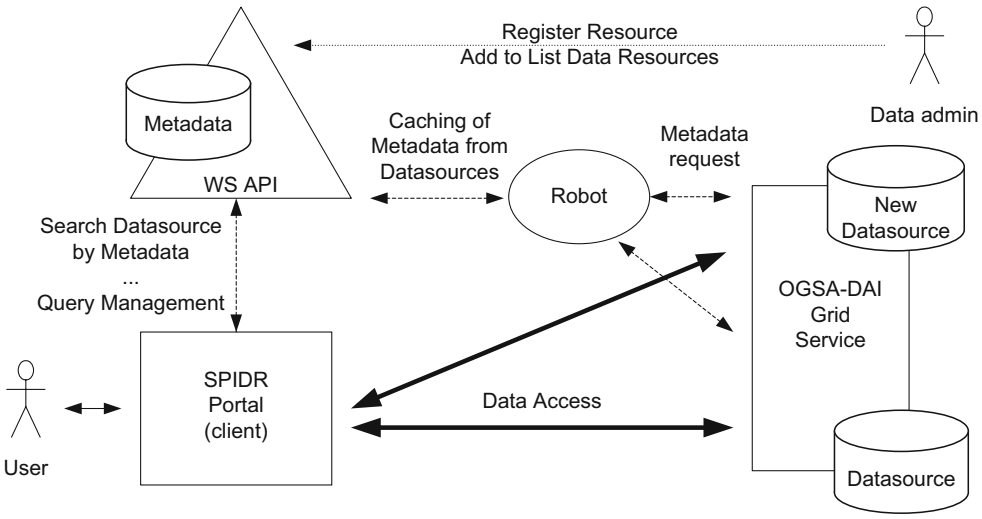

Fig. 2 A web-portal as a client for a grid of data sources

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 metadata schema (FGDC 1998), and at the same time databases with satellite telemetry can have SPASE-formatted metadata records (SPASE 2006). The SPIDR high-level metadata engine has extended search capabilities allowing it to search in specific metadata elements (keyword, title, provider, etc.) using REST web-service API. In addition, it supports Web 2.0 style[3] functionality with content versioning and moderation, direct collaborative editing of the metadata records at the SPIDR portal, a user discussion forum, e-mail and internal messaging, and wiki-style documentation and system help. SPIDR catalog level metadata search and export functions are implemented using Open Source VxOware Virtual Observatory framework (http://sourceforge.net/projects/vxoware).

The SPIDR granule-level inventory metadata registry uses an SQL database backend based built on the open-source product MySQL (MySQL 2004). 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 passed in order to update the corresponding catalog-level metadata.

Grid of virtual data sources

Web Services (WS) technology is used by SPIDR to access databases and metadata both for the SPIDR web portals (interactive interface for human users) and for the SPIDR web clients (third party programs exporting and importing data and metadata in batch mode using SOAP protocol). In addition to the WS SOAP protocol the SPIDR web portal can access databases directly using JDBC drivers. We call the JDBC-connected databases "local" and the WS-connected databases "remote" (Fig. 3). The access mode is defined in the database configuration files. If database is hosted on the server on the same local network as the SPIDR web portal, 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 over HTTP becomes the only reliable way to access the data.

SPIDR data archives are logically organized into thematic groups called viewGroups (e.g. Geomagnetic Indices, or Interplanetary Magnetic Field). Each viewGroup may include several databases or groups of database tables, which we call tables. Each table may be considered a virtual database with a single configuration file describing the access mode (local or remote) together with a URL and details required for accessing the resource.

Regarding the software architecture, each table is served by a set of Java classes for data selection, insertion and update implementing the same interface. The corresponding class names are also listed in the table configuration file. These classes are dynamically loaded by the web service

[3] Here by Web 2.0 style we mean collaborative way of on-line editing of the SPIDR metadata with a basic set of standard services including content relations, tagging, change tracking and moderation, wiki markup, user blogs, etc.

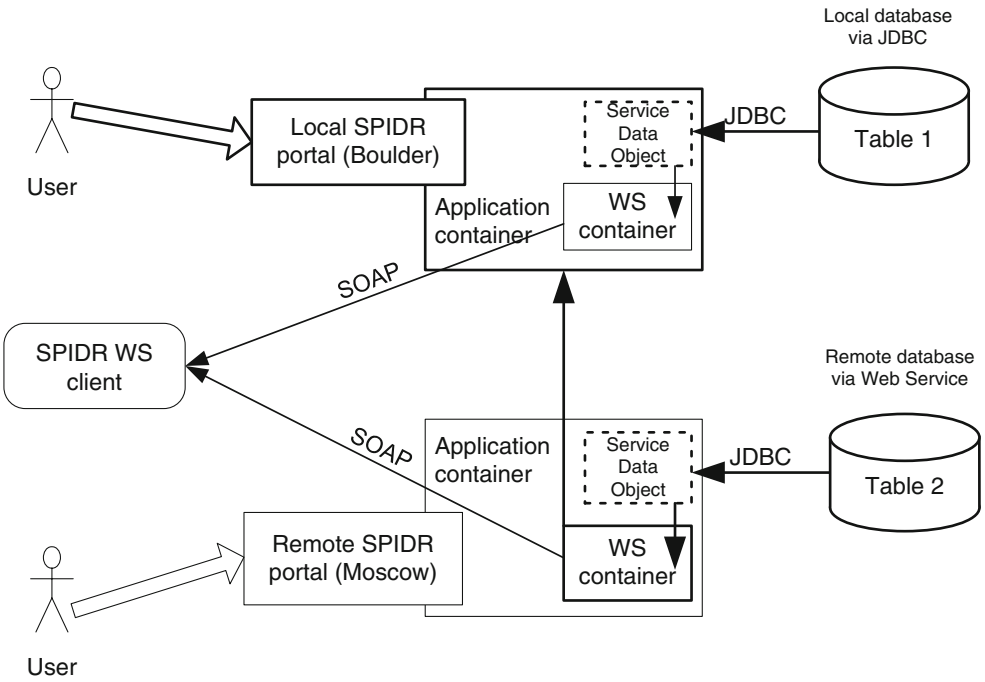

Fig. 3 Interchangeable use of local JDBC database connections and remote SOAP data source web-services in SPIDR

container and used by the SPIDR system in place of the "abstract" data access classes standing in the SPIDR core and used for any "generic" table. Each table is composed of several elements, which represent scalar observables of the space environment, such as the Bx, By and Bz (east, south, vertical) components of the interplanetary magnetic field. In many cases we also need a station (observatory or satellite) to define the scalar observable. Element values are varying in time, so within a given time interval for a given station the element defines time series which can be plotted or exported from SPIDR.

The standard procedure for adding a new database as a SPIDR data source involves implementation of the metadata manager and data export classes (we call them *classGetter* and *classMetadataUpdate*) and optional classes for parsing input and writing output data in special formats, such as WDC format for geomagnetic variations or SAO records from ionosondes.

The system currently has implemented metadata and date access web-services for the following use cases:

1. Get a metadata record for a given viewGroup (Catalog Level Metadata 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 (Inventory Level Metadata and 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).

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 (Fig. 4). The SPIDR application itself can act as a web-service and process remote calls thus allowing for chaining of the data export web services sharing the same data model in their data streams.

The central object of the CDM (Fig. 5) is the *DailyData*, which is a data container for a single physical parameter (temperature, pressure, etc.) during one observation day at a single *Station*. Besides arrays of data values, qualifiers and times (if sampling is uneven), the *DailyData* has an association with the *DataDescription*, which is the characteristic of the physical property and the data origin of the *DailyData*. Several *DailyData* objects, corresponding to the same physical parameter (particularly with the same *DataDescription*) possibly with some missing observation days can be combined in a *DataSequence* object. *DataSequence* objects, corresponding to different physical parameters, but with a common sampling value, can be combined in a *DataSequenceSet*. Attributes of the *DataSequenceSet* can be used at the presentation layer for a time series plot (plot title, logarithmic or linear *y*-axis scale, and points, lines or bars representation). Content for the *Station*, *DataDescription* and *DataSequenceSet* comes from the SPIDR metadata records.

CDM 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. In any case, a SPIDR data

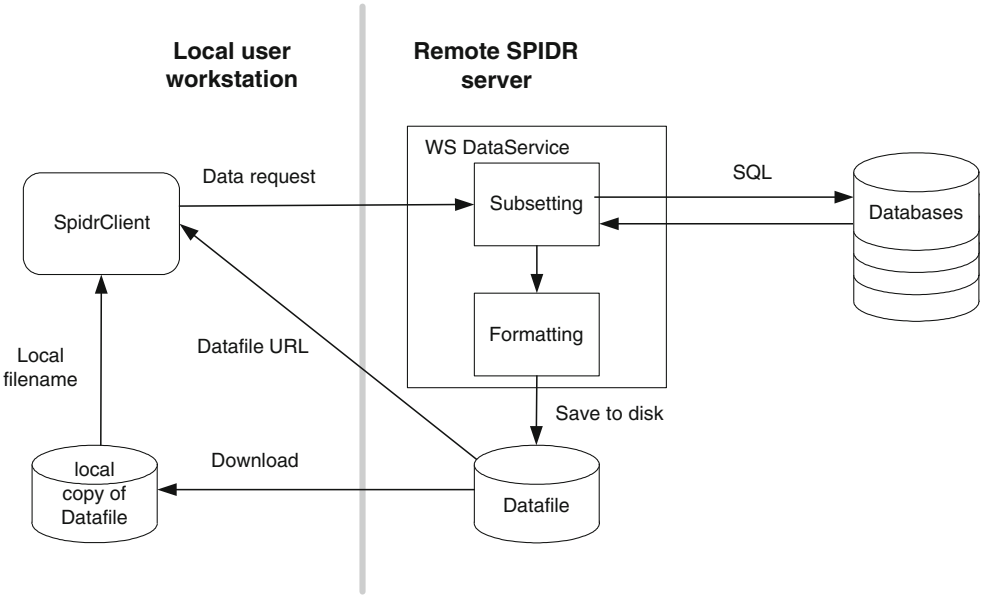

Fig. 4 Data export through the SPIDR web service

source service will supply metadata describing parameter names, units of measure, visualization options, etc., and the data accreditation describing the data origin. For geomagnetic variations the data accreditation describes the observatory which has provided the data to SPIDR.

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 Fig. 6 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 (Campbell 2001).

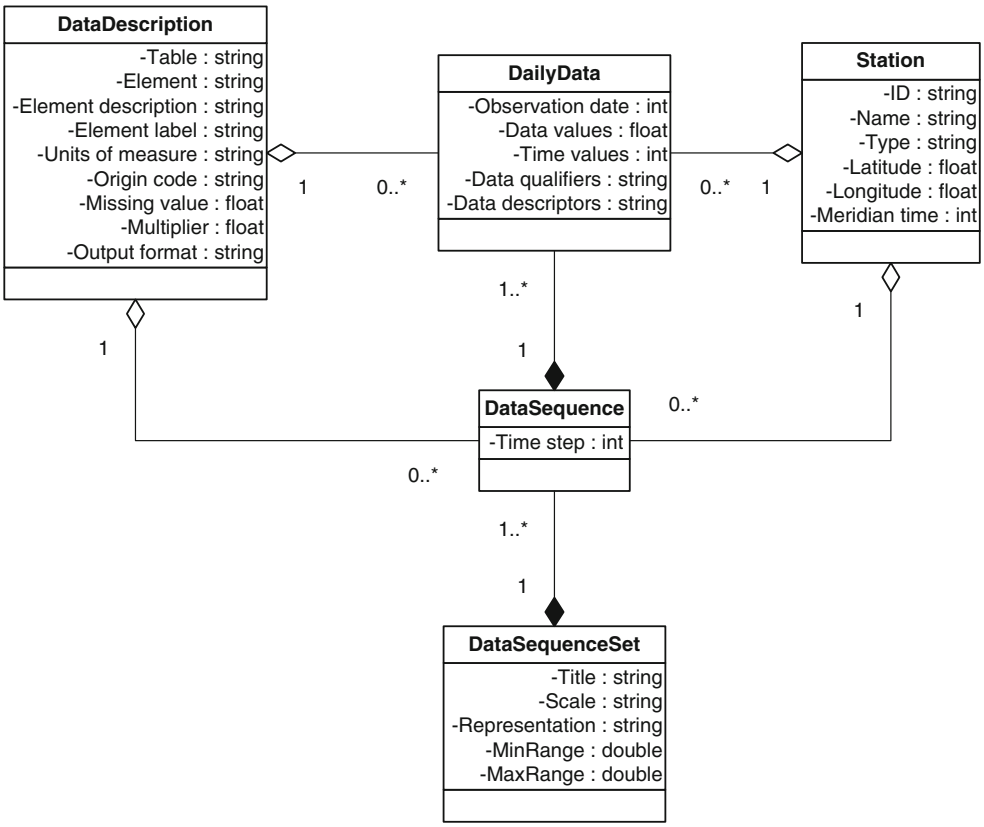

Fig. 5 SPIDR data model class diagram

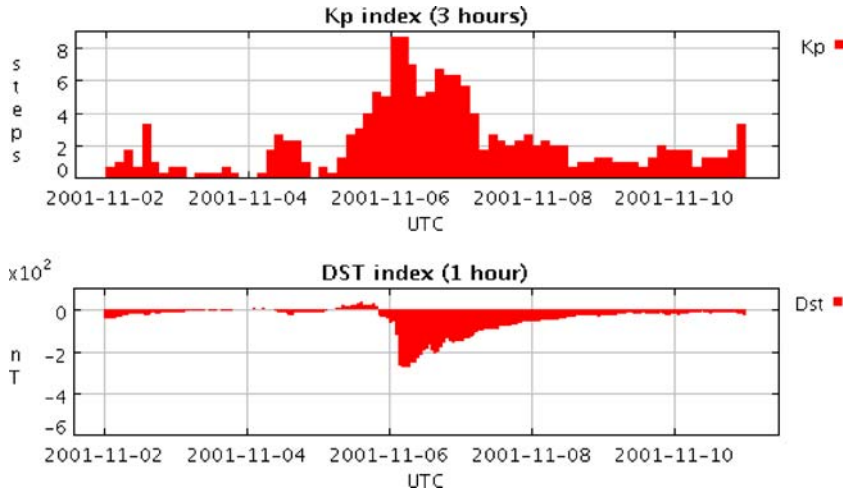

Fig. 6 Simultaneous plots of data selected from different services

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.

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.

For example, the magnetic storm shown by the time series plots in Fig. 6 is manifested in the daily solar image granule by a bright solar flare in the 164 MHz radio-telescope image. The flare erupts from a large system of sunspots visible on the solar X-ray and magnetogram images (Fig. 7). 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[4] satellite orbit; one of them is shown in Fig. 8.

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.

[4] Defense Meteorological Satellite Program, http://dmsp.ngdc.noaa.gov.

Fuzzy search web-service

SPIDR has a fuzzy search web-service which is an implementation of the Environmental Scenario Search Engine (Zhizhin et al. 2006). The web-service input is a combination of fuzzy conditions like "very low", "average", "in the range between x_1 and x_2" to be applied to some space weather parameter values; in fact it is a formalized description of any environmental event. The fuzzy search web-service is not linked to a specific type of data source. This makes the data mining extremely flexible. One can search for an event over several parameters exported by different data sources. The output of this service is a time series with values between 0 and 1 for the fuzzy likeliness of the occurrence of the event at every moment in time. Output scores that are above a user specified threshold can be used as a filter for selection of another time series.

The highest scores in the fuzzy search web-service output can be used as indications of single occurrences of the desired event. For example, a very important event in the space weather called is called a magnetic storm can be defined by the fuzzy logic expression "(low (DST)) and (high (Kp))". A SPIDR search for the year 2001 yields the event of November 6, 2001 with the Kp and DST plotted in Fig. 6. In that case meteorological satellite night-time images selected with this timing can be used for independent verification of the magnetic storm conditions in the event by showing aurora in the polar region. This is exactly the way the DMSP satellite orbits were filtered to find the clear auroral images presented in Fig. 8.

Data sink web-service

Clients can load data into SPIDR databases from files located on a local workstation, along with relevant loading

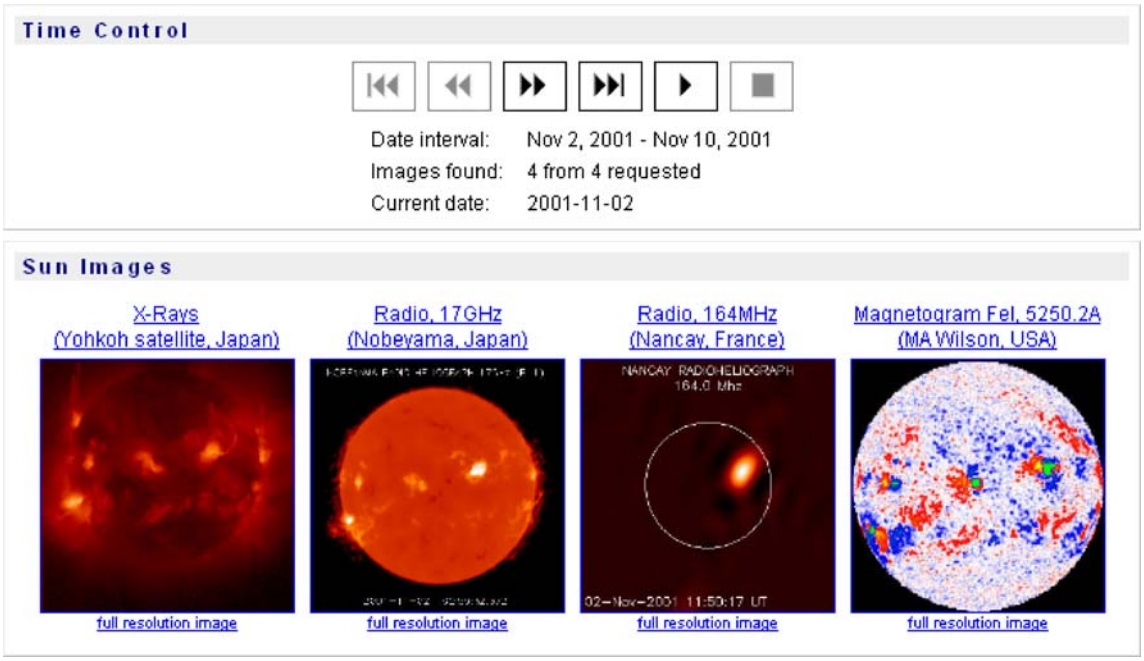

Fig. 7 SPIDR browser for solar image granules with a visible geo-efficient flare

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, update metadata inventory, 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 (Fig. 9).

Applications

Database synchronization

SPIDR databases are self-synchronizing. 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.

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.

This web-services based synchronization mechanism is a new step in automation of the data exchange between 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. The existence of several copies of the same database in a very distributed network helps support long term data preservation by protecting against local data loss. This might occur in a natural disaster.

Virtual observatory

A virtual observatory (VO) 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

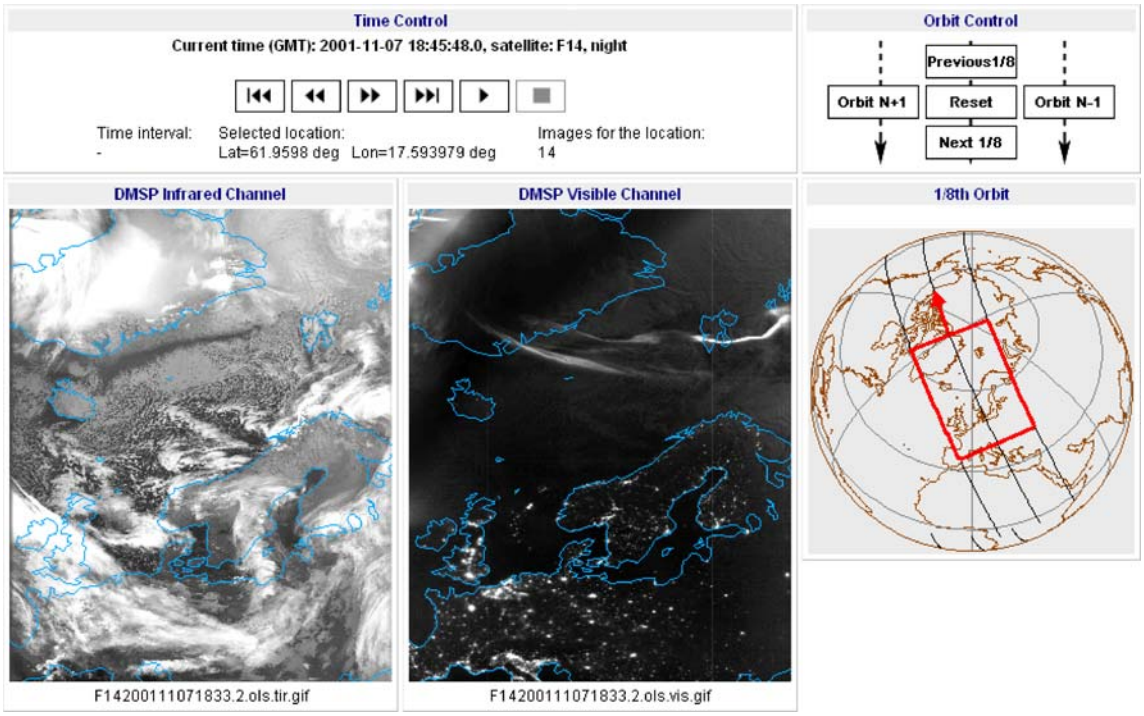

Fig. 8 SPIDR browser for DMSP image granules

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.

A VxO is a discipline specific implementation of an VO with the x representing that discipline (e.g. VMO-Virtual Magnetospheric Observatory) (Gray and Szalay 2004). We can envision that using a 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.

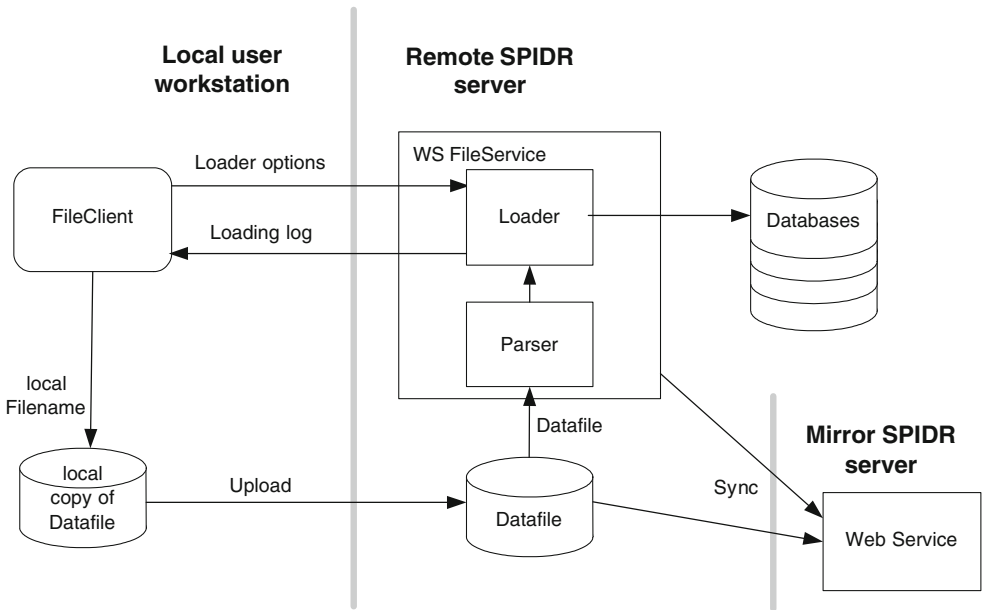

Fig. 9 Data loading and synchronization SPIDR web service

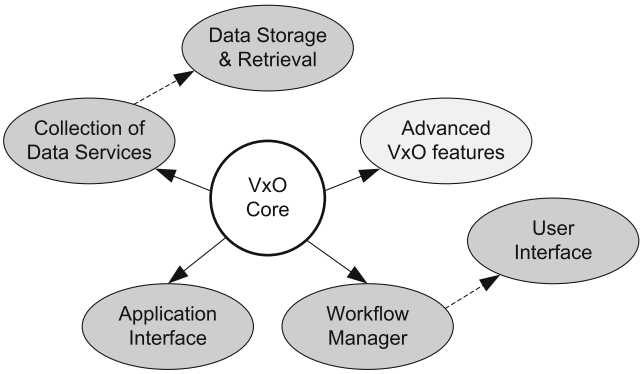

Fig. 10 Typical virtual observatory services

- 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.

The SPIDR system as implemented includes most of the elements of a VO including catalog level metadata service. Catalog level metadata follows the FGDC and SPASE standards. In addition, the metadata services within SPIDR have the inventory level (available observation dates) and more detailed metadata (station description and data manager reference) objects.

As shown in Fig. 10, a virtual observatory may consist of a number of services built around a community based core. Because SPIDR is built around the Grid core it is easily adaptable to support a VO 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, http://virbo.org/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. ViRBO metadata repository can import catalog level metadata from SPIDR. ViRBO users can select radiation belts related data with the SPIDR Data Source WS.

A list of requirements to the modern VO workflow manager component will include

- Support of many possible levels of user interaction
- Support of user community-centric views including
  - Community newsfeeds
  - Portals to other VxO's
  - Community specific on-line tutorials and document repositories
  - Forum, wiki and blog capabilities
  - Software libraries and downloadable tools.

The SPIDR user interface implements all of these criteria by using the Apache Struts[5] (http://struts.apache.org) framework to define workflow. This 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.

Another element to a VO is the Application Programming Interface (API), here the system must provide:

- Web-services based data flow and data transformations;
- Metadata search and data export API to interface with other VxO's;
- Common computational and data models;
- Mechanisms for creating derived data products;
- Markup for events of interest, e.g. magnetic storm detection.

Obviously many of these items map directly to the Grid paradigm providing a strong link between a VO and 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.

Space weather reanalysis

One of the more important projects enabled by the SPIDR system is the Space Weather Reanalysis (SWR) (Ridley and Kihn 2004). The objective of the SWR was to generate a two solar cycle (22 years) space weather representation using physically consistent data-driven space weather models. This was accomplished by using data made accessible through SPIDR and models integrated through Grid technologies. The resulting product is an enhanced look at the space environment on consistent grids, time resolution, coordinate systems and containing key fields allowing a scientist to quickly and easily incorporate the impact of the near-Earth space climate in models which can ingest this data. Before this project there were no climatological archives for the space-weather environment. Just as with terrestrial weather it is crucial for scientists to understand both daily weather forecasts as well as long term climate changes which affect operations.

The results of this project support further tools for intelligent data mining, classification and event detection which are applied to the historical space-weather database. This provides a reasonable starting point for the user interested in modeling the effect of the near-earth space environment on society.

[5] Apache Struts is a free open-source framework for creating dynamic web content with JavaServer pages. Struts can interact with databases and business logic engines to customize a response and to control the application workflow.

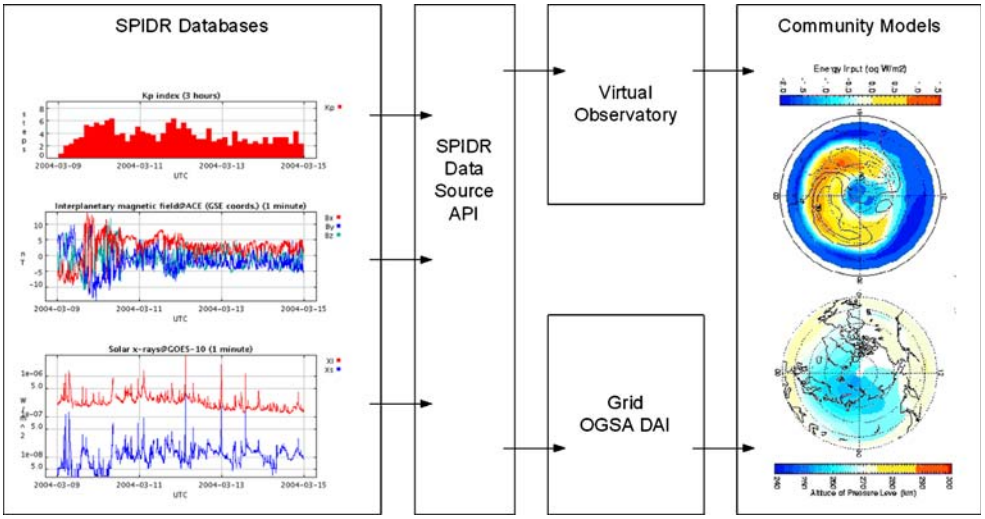

Fig. 11 SPIDR services used for the space weather reanalysis

As shown in Fig. 11 the SPIDR system provides integrated data access to a suite of models both through community specific interfaces and through the Grid service layer. By removing the need to separately manage the data at each provider the SPIDR systems enables developers to concentrate more directly on the science and integration issues associated with the models. In the case of the SWR four models were linked which comprise a reasonable near-Earth space environment. All of the data necessary for these models was accessed from WDC holdings using SPIDR Data Source WS. This use case is typical for the environmental sciences, where the data is often distributed, in diverse formats and difficult to match for a given combination of time, parameters and geographic location. The service nature of a Grid provides a convenient infrastructure for search and accessing the necessary resources to meet a given scientific objective.

Future work on grid data mining

In a June 1999 Nature article entitled "It's sink or swim as a tidal wave of data approaches. ... Are scientists ready for the flood?" the authors stated "Most researchers are accustomed to studying a relatively small data set for a long time, using statistical models to tease out patterns. At some fundamental level that paradigm has broken down." (Reichhardt 1999) This is more true today than ever, no longer are scientists restricted to a small collection of data that they themselves collect and manage, rather they have the possibility to interact with literally petabytes of environmental archives. Many of the important discoveries are coming not from a single discipline but from cross-discipline work where the integration of heretofore separate archives yields new and important discoveries.

In this arena the Grid plays an important part, because faced with an incredible volume of data in diverse formats, archives and types the scientists need a tool kit to point them to the relevant bits and maximize scientific productivity. An important tool which can be brought to bear on this is data mining. Data mining is the act of discovering by autonomous mathematical algorithms heretofore unknown relations between data. This is often used in the business community to match consumer buying patterns to goods and advertising for example. In the environmental sciences data mining has several important applications these include data quality control, human linguistic translation, event and trend detection, data classification and forecast, and deviation detection.

In particular deviation detection has become an important topic with items like global warming, hurricane intensity, solar output changes all impacting the human environmental system. Grids enable environmental data mining by:

1. Making access to cross disciplinary data transparent to the user.
2. Providing access to services necessary to prepare the data (e.g. sub setting, transforms).
3. Integrating data mining services.

Unfortunately, Grids do not assist with the all important data validation and verification that is required before data mining tools can do effective work.

SPIDR data services can be used as an Open Grid Services Architecture—Data Access and Integration (OGSA-DAI, http://www.ogsadai.org.uk) resource using a data resource plugin mechanism. OGSA-DAI is a middle-

[6] The Enabling Grids for E-sciencE (EGEE) project brings together computing resources and researchers from 240 institutions in 45 countries to provide a seamless Grid infrastructure for e-Science that is available to scientists 24 hours-a-day. The EGEE project is funded by the European Commission.

ware product which supports the representation of various data resources, such as relational or XML databases, to the Internet and Grids (Antonioletti et al. 2005). It can be used as a data service layer both in the Globus Toolkit 4 and soon in the gLite grid implementations. That makes it possible to build grid portals and to run the space weather models within EGEE[6] infrastructure (http://www.eu-egee.org) using SPIDR as an embedded (meta)data web service.

For example, the OGSA-DAI Grid data services can use Environmental Scenario Search Engine (ESSE, http://esse.wdcb.ru) for data mining in SPIDR. The ESSE user interface can translate natural language descriptions of environmental events (i.e. "large magnetic storm") without having to assign specific values to parameters involved (Zhizhin et al. 2006): using the ESSE engine, the "large magnetic storm" event can be described as a fuzzy logic query to search for a combination of "low DST and large Kp" values of the global geomagnetic indices stored in SPIDR databases.

The data held in SPIDR can be quality controlled by peer-matching techniques where stations are compared to nearest neighbors to see if the observations are "similar" (in a fuzzy mathematical sense). This quality control technique was used for the data export from SPIDR to feed the numerical space weather models in the NGDC Space Weather Reanalysis. To data mine the Space Weather Reanalysis products, we were able to use similar analog forecast technique for ionospheric potentials (Kihn et al. 2006).

Conclusions

It is our belief that increasing data volumes demand application of new tools and methods to maximize scientific efficiency. It is out belief that software tools and mathematical methods exist which, provide analysis, classification, access, discovery and forecast methods for large volume data sets. The 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

Antonioletti M, Atkinson MP, Baxter R, Borley A, Chue Hong NP, Collins B, Hardman N, Hume A, Knox A, Jackson M, Krause A, Laws S, Magowan J, Paton NW, Pearson D, Sugden T, Watson P, Westhead M (2005) The design and implementation of Grid Database Services in OGSA-DAI. Concurrency Comput Prac Ex 17(2–4):357–376

Barkstrom BR, Hinke TH, Gavali S, Smith W, Seufzer WJ, Hu C, Cordner DE (2003) Distributed generation of NASA Earth Science Data Products. Journal of Grid Computing 1:101–116. Online at http://www.nas.nasa.gov/News/Techreports/2005/PDF/nas-05-006.pdf

Campbell W (2001) Earth magnetism: a guided tour through magnetic fields. Academic, San Diego, p 151

Domenico B, Caron J, Davis E, Kambic R, Nativi S (2002) Thematic Real-time Environmental Distributed Data Services (THREDDS): incorporating interactive analysis tools into NSDL. J Dig Inform 2 (4):114. 2002-05-29. Online at http://jodi.tamu.edu/Articles/v02/i04/Domenico/

FGDC Federal Geographic Data Committee (1998) Content standard of digital geospatial metadata, Version2, 1998. Online at http://www.fgdc.gov/metadata/contstan.html

Fielding RT (2000) Architectural styles and the design of network-based software architectures. Doctoral dissertation. University of California, Irvine. Online at http://www.ics.uci.edu/~fielding/pubs/dissertation/fielding_dissertation.pdf

Foster I, Kesselman C, Tuecke S (2001) The anatomy of the grid: enabling scalable virtual organizations. Int J High Perform Comput Appl 15(3):200–222. doi:10.1177/109434200101500302. Online at http://www.globus.org/alliance/publications/papers/anatomy.pdf

Gallagher J, Potter N, Sgouros T (2007) DAP data model specification. Online at http://www.opendap.org/pdf/dap_objects.pdf

GCMD (2008) NASA global change master directory. Online at http://gcmd.nasa.gov

Gray J, Szalay A (2004) Where the rubber meets the sky: bridging the gap between databases and science. Microsoft Research Technical Report MSR-TR-2004-110. Microsoft Research Redmond, WA. Online at ftp://ftp.research.microsoft.com/pub/tr/TR-2004-110.pdf

ICSU (1996) Guide to the World Data Center System. International Council of Scientific Unions, Paris, France. Available via http://www.ngdc.noaa.gov/wdc/guide/wdcguide.pdf

Kihn EA, Zhizhin M, Siquig R, Redmon R (2004) The Environmental Scenario Generator (ESG): a distributed environmental data archive analysis tool. CODATA Data Sci J 3:10–28. doi:10.2481/dsj.3.10. Online at http://www.jstage.jst.go.jp/article/dsj/3/0/10/_pdf

Kihn E, Zhizhin M, Kamide Y (2006) An analog forecast model for the high-latitude ionospheric potential based on assimilative mapping of ionospheric electrodynamics archives. Space Weather 4:S05001. doi:10.1029/2005SW000199

Loughran S, Smith E (2005) Rethinking the Java SOAP Stack, Technical Report HPL-2005-83 20050517, Hewlett-Packard. Online at http://www.hpl.hp.com/techreports/2005/HPL-2005-83.pdf

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

MySQL AB (2004) MySQL Reference Manual. Online at http://www.mysql.com/doc/en/index.html

Nieto-Santisteban M, Szalay A, Thakar A, O'Mullane W, Gray J, Annis J (2004) When Database Systems Meet the Grid. Microsoft Research Technical Report MSR-TR-2004-81, Microsoft Research Redmond, WA. Online at ftp://ftp.research.microsoft.com/pub/tr/TR-2004-81.pdf

Reichhardt T (1999) It's sink or swim as a tidal wave of data approaches. Nature 399(6736):517–520 10 June

Ridley AJ, Kihn EA (2004) Polar cap index comparisons with AMIE cross polar cap potential, electric field, and polar cap area. Geophys Res Lett 31:L07801. doi:10.1029/2003GL019113

Siquig R, Lowe S (1996) The master environmental library: an environmental data source for DoD applications. Defense Technical Information Center ADA335410. Online at http://handle.dtic.mil/100.2/ADA335410

SPASE (2006) Space Physics Archive Search and Extract (SPASE) data model standard. Online at http://www.spase-group.org/data/

Zhao Y, Wilde M, Foster I, Voeckler J, Dobson J, Glibert E, Jordan T, Quigg E (2006) Virtual data Grid middleware services for data-intensive science, Concurrency and Computation: Practice & Experience, vol. 18, pp. 595–608. Online at http://www.griphyn.org/documents/document_server/uploaded_documents/doc-1540-cpe-zhao-crc.pdf

Zhizhin M, Poyda A, Mishin D, Medvedev D, Kihn E, Lyutsarev V (2006) Scenario Search on the Grid of Environmental Data Sources. Microsoft Research Technical Report. Online at http://research.microsoft.com/research/pubs/view.aspx?type=Technical%20Report&id=1116