Introducing Subset.Org: A Portal for Subsetting Applications and Related Tools

Danny Hardin, Matt He, and Sara Graves

Information Technology and Systems Center[1]
Department of Atmospheric Science[2]
University of Alabama in Huntsville
Huntsville, AL - 35899

1. INTRODUCTION

There is currently a tremendous volume of scientific data available to researchers. Numerous Earth orbiting satellites, aircraft and in-situ instruments gather massive quantities of information daily. The recently launched Aqua Satellite with its suite of instruments will introduce another Terabyte of data into the pool of Earth Science data daily.

Researchers who are not studying global phenomena are faced with the problem of obtaining subsets from this trove of information that apply to their specific region and/or time period. Fortunately there are many applications that can extract specific subsets of the data by region, time or event. These applications and the problems they solve vary according to the manner in which the data are formatted and to some extent the geographic regions covered. Polar regions have their own special geolocation problems.

The Information Technology and Systems Center at the University of Alabama in Huntsville has developed a subset portal, located on the web at www.subset.org, to help data users find and use the appropriate subsetting tools for a particular job. Subset.org provides a common location for subsetting information that helps users rapidly sift through an inventory of available subsetting tools, including data mining applications and data fomat information. It also provides a means for user information interchange and discussions. A personalized portal page allows users to interactively manage a group of subsetting services to fit their individual needs.

2. SUBSETTING APPLICATIONS

Data volumes can be reduced and scientific investigation focused through the use of tools that extract only the relevant portion of a data set. Subsetting a global data set over a specific time interval or geographic location are common uses for subsetting [1]. However, subsetting applications exist that are able to extract subsamples of a data set or in some cases a data mining operation can be employed to search through a data set in search of specific phenomena. Subset.org lists a number of these applications [2]. A sampling of major applications follows.

2.1 The Advanced Very High Resolution Radiometer (AVHRR) Web Subsetting Tool (The Earth System Science Workbench, The University of California Santa Barbara)

The Advanced Very high Resolution Radiometer (AVHRR) is a broad-band, four or five channel scanner, sensing in the visible, near-infrared, and thermal infrared portions of the electromagnetic spectrum. This sensor is carried on the National Oceanic and Atmospheric Administration's (NOAA's) Polar Orbiting Environmental Satellites (POES), beginning with TIROS-N in 1978.

The AVHRR Product generator is an application designed to improve the process of requesting and receiving data from a database of L1B images. The process involves four easy steps: temporal search, profile creation/search, product request, and retrieval of the finished product. The database of L1B images now contains AVHRR data for the West Coast of North America from June 1999 to present. Earlier AVHRR data back to October 1993 can be retrieved upon request from archives.

Anyone with access to the web can use a web browser to search the metadata and request a spatial and/or spectral subset of an image. The image is then delivered to the user in the requested format via an email message that includes a universal resource locator (url), or Web address, from which the image can be viewed and/or downloaded.

2.2 AVHRR Oceans Pathfinder Subsetter System (Physical Oceanography Distributed Active Archive Center, Jet Propulsion Laboratory)

The NOAA/NASA AVHRR Oceans Pathfinder sea surface temperature data are derived from the 5-channel Advanced Very High Resolution Radiometers (AVHRR) on board the NOAA -7, 9, 11 and 14 polar orbiting satellites. Daily, 8-day and monthly averaged data for both the ascending pass (daytime) and descending pass (nighttime) are available on equal-angle grids of 4096 pixels/360 degrees (nominally referred to as the 9km resolution), 2048

pixels/360 degrees (nominally referred to as the 18km resolution), and 720 pixels/360 degrees (nominally referred to as the 54km resolution or 0.5 degree resolution). Global files are available through PO.DAAC ftp or order form and desired regions are available through the AVHRR Pathfinder subsetting system.

2.3 Coarse-Grain SSM/I Subsetting (Information Technology and Systems Center)

A specialized subsetter for Special Sensor Microwave/Imager (SSM/I) data is currently available. Each SSM/I data file is composed of a set of geographically discrete "subgranules". This subsetter very quickly extracts the subgranules that cross the uer's area of interest.

2.4 General-Purpose Subsetting (Information Technology and Systems Center)

This general-purpose subsetter provides general-purpose subsetting services using specialized data readers and writers with general-purpose analysis routines. This provides a framework that can be used for many types of data processing, including subsetting, subsampling, averaging, and format conversions. Currently many common data types and formats are supported.

2.5 Graphical Interface for Subsetting, Mapping, and Ordering (GISMO) (National Snow and Ice Data Center)

The Graphical Interface for Subsetting, Mapping, and Ordering is a Web-based search, order, and subsetting interface for gridded data at NSIDC. GISMO allows users to search data sets by collection, parameter, and date. Users can also specify an area of interest and spatially subset data to reduce the total volume of delivered data. GISMO currently provides search and order capabilities for AVHRR 1.25-km and 5-km Northern and Southern Hemisphere EASE-Grids; SMMR 25-km Northern, Southern, and Global EASE-Grids; SSM/I 12.5- km and 25-km Northern, Southern, and Global EASE-Grids; and TOVS 100-km Northern Hemisphere EASE-Grids. GISMO subsets the data and automatically stages them to an anonymous ftp site for user access.

2.6 HEW: A Dataset-Independent Subsetter for HDF-EOS Files (Information Technology and Systems Center)

HDF-EOS is the preferred format for the storage of data in NASA's Earth Observing System Data and Information System (EOSDIS) [3]. HDF-EOS expands the capabilities of HDF by adding *swath*, *grid*, and *point*. data types that are specifically designed for the storage of geophysical data.

HEW (HDF-EOS Web-based subsetter) can extract a subset of any grid or swath data file that is in HDF-EOS format. Subsetting can be performed on latitude and longitude (rectangular areas), date and time span (swath data) or by dataset parameter, e.g., instrument or sensor

HEW is also capable of subsampling by extracting every Nth point of data. As a stand-alone subsetter, HEW uses a web-based front-end to gather the user's subsetting criteria and then submits the subsetting job to a batch queue.

A companion program to HEW, entitled SPOT, can be used to check HDF-EOS files for subsettability. SPOT is invoked using a simple command-line interface. It checks to determine if the file exists and is readable, the file is in HDF format, the file is in HDF-EOS format, and if the file contains valid HDF-EOS structures

SPOT can be invoked in "verbose" or "silent" mode. In "verbose" mode, SPOT displays the HDF-EOS structures as it checks them. This can be used to view the structure of an unknown HDF-EOS file. In "silent" mode, only error messages are output. In both cases, the exit status is set to reflect the "subsettability" of the file.

2.7 On-Demand Subsetting: The PM-ESIP (Information Technology and Systems Center)

The Passive Microwave Earth Science Information Partner project (PM-ESIP) provides science researchers and users with the capability to interactively customize and retrieve hydrologic datasets for use in process studies and regional and global climate studies.

The PM-ESIP provides web-based access to interactive tools that can subset pre-generated datasets, generate new datasets on demand from lower-level global or regional data, extract coincident data from multiple datasets, and convert from swath to a variety of gridded map projections.

2.8 Polar Spatial Query (PSQ) (National Snow and Ice Data Center)

The Polar Spatial Query (PSQ) tool is a web-based search, order, and subsetting interface for swath and scene data at NSIDC. PSQ allows users to search for orbit and scene data -- currently SSM/I antenna temperatures and AVHRR Polar 1 Km Level 1B Data -- by collection, parameter (channel), date, and region of interest.

The main advantage of the PSQ is the ability to search for orbit and scene data using an orbit model, yielding search results that are much more accurate. Additionally data are mapped to a common grid covering the user's region of interest.

PSQ regrids the data and automatically stages them to an anonymous ftp site for user access.

2.9 The Subsetter/Format Converter (SFC) (Information Technology and Systems Center)

The SFC is a specialized application built for different end users using customized combinations of readers, operations, and writers. It utilizes selected readers and writers for data sets and the generic subsetting/subsampling modules. The SFC deciphers metadata information for the supported data sets providing end users different choices. The SFC was built using Java, providing platform independence ease of use. The SFC resides on the end user's/scientist's workstation and works on local files. It offers the capability to subset data files based on spatial bounds, temporal extent, different data fields, data subsampling, or any combination of these parameters.

Figure 1. The index page of Subset.org contains a predefined set of portlets (shown with yellow borders.

3. SUBSET.ORG PORTAL TECHNOLOGY

The ultimate goal of the Subset.org portal is to provide access to resources that contribute to the increased use of Earth science data, including applications, tutorials, documentation, community specific news and so forth. The Subset.org portal acts like a central hub where information from multiple sources is made available in an easy to use manner even though the original sources of the information may be widely distributed.

The portal is constructed from numerous building blocks known as portlets. These portlets contain information on specific subjects. The subset.org home page contains several portlets that appear as separate rectangular regions with yellow borders (Figure 1). The ITSC web team controls the structure of this page. But, users may subscribe to Subset.org and upon registration create their own customized layout selecting portlet types that suit their needs (Figure 2).

Figure 2. Personalized portal page for a registered user.

Six portlet types are defined for Subset.org each with content expected to be beneficial to the user community. The portal features numerous instances of these portlet types.

1) Subsetting Applications
The main portlet type of Subset.org provides information on available subsetting applications and related tools such as those described in section two.

2) Documentation
The documentation portlets contain reference information such as tutorials, examples, user manuals and other documentation.

3) News
News portlets receive information from news providers using RSS technology. Available news providers include selected Federation of Earth Science Information Partners, subsetting application providers, and major news agencies. Registered users can customize the news section by selecting news source and subject topics that match their interests. News portlets are automatically refreshed whenever a new news item is available.

4) Community
These portlets allow users to communicate with each other, participate in subset portal events, and facilitate portal expansion by providing a means to inform the community of new subsetting applications.

5) Administration
The administration portlets allow the ITSC web team to maintain the site easily and efficiently.

6) General Information
These portlets give registered users a wealth of choices for creating custom portals. Information such as stock portfolio, weather, favorite bookmarks, calendars, etc. can be added as the user desires.

4. IMPLEMENTATION TECHNOLOGIES

The Subset.org portal was built using Jetspeed from the Jakarta Apache Project Group. Jetspeed utilizes Java Servlet and JavaServer Page technologies. It also uses XML, RSS or

SMTP as the means for presenting content. Tools for portal access via a web browser or wireless devices are also included. Other Jetspeed technologies include Jserv and Velocity as the servlet engine, Xerces and Xalan for XML parsing and styling, Castor for data marshalling, Turbine for the application framework and Cocoon for generating pages from XML [4, 5, 6]. Although Jetspeed is still in the development phase, it's already gained a lot of attention from many portal developers.

Jetspeed is designed to be a gateway to many information sources, including resources from the Internet as well as custom-coded access to services within an organization. XML is used within Jetspeed as the mediating format for incoming data. Subset.org uses Jetspeed's architecture and source code as the basic components although many new portlets have been added. The key features in Jetspeed include the following [4]:

- Java Portlet API specifications for creating new portlets
- Template-based layouts with JSP and Velocity
- Remote XML content feeds via Open Content Syndication
- Custom default home page configuration
- Database user authentication
- In-memory cache for quick page rendering
- Rich Site Summary support for syndicated content
- Integration with Cocoon, WebMacro and Velocity so that users can develop with the newest XML/XSL technology.
- Wireless Markup Language (WML) support
- XML based configuration registry for portlets
- Full Web Application Archive (WAR) support
- Web Application development infrastructure
- Local caching of remote content
- Synchronization with Avantgo
- Portability across all platforms that support JDK 1.2 and Servlet 2.2
- Integration with Turbine modules and services
- Profiler Service to access portal pages
- Persistence Service so all portlets can easily store state per user, page and portlet
- Skins so that users can choose colors and display attributes
- Customizer for selecting portlets and defining page layouts
- Database storable PSML
- Administration via Jetspeed security portlets by user, group, role and permission
- Role-based portlet security access

4.1 RDF Site Summary (RSS)

RDF (Rich Data Format) Site Summary (RSS) is one of the most widely used XML formats on the Web [7]. RDF Site Summary (RSS) files provide an open method for syndicating and aggregating web content. Using RSS files, websites can create a data feed that contain logos, links, headlines, and summaries. Other websites can incorporate this information into their pages automatically. These techniques provide users with up-to-date information.

The RSS format originated with the sites *My Netscape* and *My UserLand*, both of which aggregate content derived from XML news feeds. Because it is one of the simplest XML applications, RSS became popular among developers who needed to perform similar tasks.

Subset.org relies heavily on RSS feed to generate the content of most portlets. RSS syndication is not only used in presenting news items, but also to provide information on subset applications, documentation, software, tutorials, etc. The advantage of this approach is: the RSS feed source is located at different websites and can be easily modified by the ITSC web team to provide up to date information to the portal without changing the portal itself. An example RSS file is shown in Figure 3. This is the news syndication file from www.esipfed.org.

```
<?xml version="1.0" encoding="ISO-
8859-1" ?>
<rss version="0.92">
<channel>
<title>ESIP Federation News</title>
<link>http://www.esipfed.org/news/in
dex.jsp</link>
<language>en</language>
<lastBuildDate>09/06/02/14:30CEST
</lastBuildDate>
</channel>
<item>
<title>TEACHING OLD DATA NEW
TRICKS</title>
<link>http://www.esipfed.org/news/ne
wsitems/02-8_old_data.html</link>
</item>
<item>
<title>NASA SATELLITES HELP
HURRICANE FORECASTERS SINCE 1992's
DESTRUCTIVE HURRICANE ANDREW</title>
<link>http://www.esipfed.org/news/ne
wsitems/02-
8help_hurricanes.html</link>
</item>
<item>
<title>New Items</title>
<link>http://www.esipfed.org/news/in
dex.jsp</link>
</item>
<textinput>
<title>Search News on
ESIPFED.ORG</title>
```

```
<description>Search Earth Science
stories</description>
<name>keyword</name>
<link>http://www.esipfed.org/news/se
arch.jsp</link>
</textinput>
</rss>
```

Figure 3, News syndication file from www.esipfed.org

4.2 Portlets

Jetspeed provides a framework that allows extensions – called portlets - to be plugged into the portal. Similar to a servlet in an application within a web server, a portlet is an application within the portal. Portlets are the building blocks used to construct personalized portal pages too. Jetspeed provides seven basic types of portlets:

- HTML Portlet, which is used to render a file containing HTML tags
- JSP Portlet, which is used to render a Java Server Page
- RSS Portlet, which is used to parse the RSS feed and render the RSS content
- Velocity Portlet, which is used to render a Velocity page (similar to JSP)
- Web Page Portlet, which is used to render a complete HTML web page
- XSL Portlet, which transforms and renders an XML document
- Database Browser Portlet, which provides database access and renders the query results

Similar to Servlet APIs, Jetspeed has Portlet APIs. With portlet APIs, a portal developer can build a plug-in portlet and program it to do anything required to perform a particular task. Jetspeed's AbstractPortlet class provides foundations for all portlets. A new portlet can be developed by extending the AbstractPorlet class. Figure 4 shows a simple Portlet example. To make a portlet perform a particular task, the setTemplate method and getContent method need to be overwritten. Subset.org utilizes all Jetspeed portlets and some newly developed ones.

```
package com.subset.portal.portlets;
import
org.apache.jetspeed.portal.portlets.
AbstractPortlet;
import
org.apache.turbine.util.RunData;
import
org.apache.ecs.ConcreteElement;
import org.apache.ecs.StringElement;

public class HelloWorldPortlet
extends AbstractPortlet
{
    public ConcreteElement
getContent (RunData runData)
    {
        return (new StringElement
("Hello World!"));
    }
}
```

Figure 4. Example of portlet programming.

4.3 Mailing Lists and Mailing Archive

To promote interaction within the Earth science community, a mailing list has been created and a searchable mailing archive has been configured. Users are encouraged to use this mailing list to exchange information within the community.

The web team chose Mailman as the software for mailing and archiving. Mailman is open source software and has been widely used in numerous websites.

5. USING THE PORTAL

5.1 User Configurable Features

Subset.org uses Jetspeed's Portal Structure Markup Language (PSML) for page content description. Each portal page is a PSML file, which contains all the portlets to be rendered and the positions specified for each. The default portal page contains a predetermined set of portlets. One of them is the Logon portlet. When a user visits the website, he/she can register allowing the customization of his/her own Subset.org portal. Once a new user account is created, Jetspeed automatically creates a new directory in the server file system and generates a new PSML file using the default settings. Then users can customize the portal page by selecting different portlets and layout. Figure 5 shows some of the portlets provided by Jetspeed. The content of each portlet can also be changed if that portlet allows content reconfiguration. Changes can be made by clicking the control icons located on the upper right corner of the portlet (as shown in Figure 6). In this way, a user can select the information he/she wants to see on the personalized portal page. The new settings will be saved in the PSML file so that the same content will be rendered on subsequent activations of the page.

Figure 5. Users can customize their portal by selecting from a list of predefined portlets.

Users can not only control the content of the portal page, but they can also change the layout of the portlets. Several different layouts can be selected and more layouts can be implemented if desired. For example, as seen in Figure 1, the Portlets are presented as a 3-column page. Users can change that into a 2-column page or a single column. The color scheme of each portlet can also be changed.

5.2 Posting New Subsetting Items

Users are encouraged to present new subsetting tools and information to the subset community. One way is through the subset mailing list; another way is by utilizing special portlets that support on-line submission of information. Similar to other portlets, submitting portlets contain links by which a submitting process can be performed. However, all new Subsetting applications have to be reviewed before they appear on the subset portal page. This ensures that improper content is not posted.

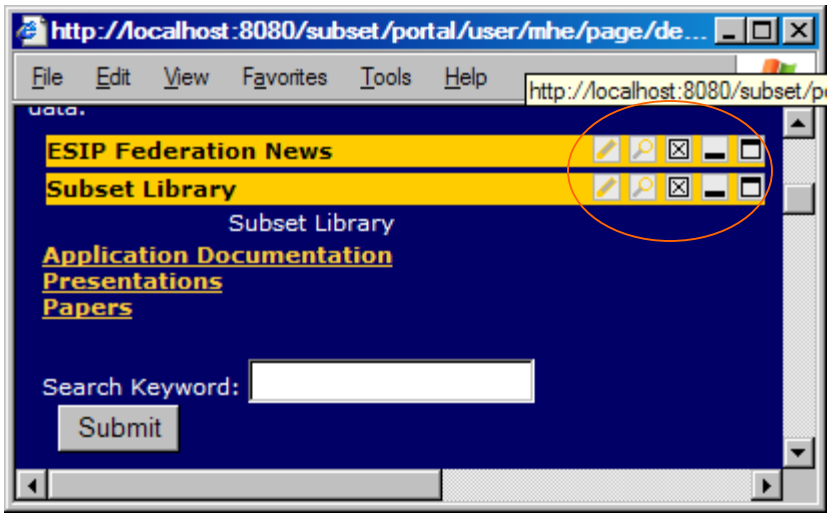

Figure 6. Portlets can be configured by using controls in the upper right corner.

Reference:

1. "A Dataset-Independent Subsetting Prototype", Matt Smith, Bruce Beaumont, Sara J. Graves, HDF-EOS Vendors Workshop, Goddard Space Flight Center, September 8-10, 1997

2. "The Role of Data Mining in Earth Science Data Interoperability", Rahul Ramachandran, Helen Conover, Sara J. Graves, Ken Keiser, John Rushing, ASPRS Annual Conference, Conference on Remote Sensing Education (CORSE), Education for the Next Millennium, Portland, Oregon, May 17-22, 1999

3. "Subsetting Data for EOSDIS", Matt Smith, Bruce Beaumont, Susan McCoy, Sara J. Graves, American Meteorological Society 13th International Conference on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology, February 2-7, 1997

4. Jakarta Project, http://jakarta.apache.org/jetspeed/site/index.html

5. Jeff Linwood, Build portals with Jetspeed, Java World, July 2001. http://www.javaworld.com/javaworld/jw-07-2001/jw-0727-jetspeed_p.html

6. Edd Dumbill, XML at Jetspeed, May, 2000 http://www.xml.com/pub/a/2000/05/15/jetspeed/

7. James Lewin, An introduction to RSS news feeds, http://www-106.ibm.com/developerworks/library/w-rss.html