Design Of A Webservice and Application For Quick Easy Access To Subsets Of Petabytes Of Air Quality Data

Todd Plessel[1], Matt Freeman[1], and Jim Szykman[2]

[1]Lockheed Martin, Environmental Modeling and Visualization Laboratory, National Computing Center, Research Triangle Park (RTP), NC 27711, USA

[2]Environmental Sciences Division, National Exposure Research Laboratory, Office of Research and Development, USEPA, c/o NASA Langley Research Center, Hampton, VA 23681, USA

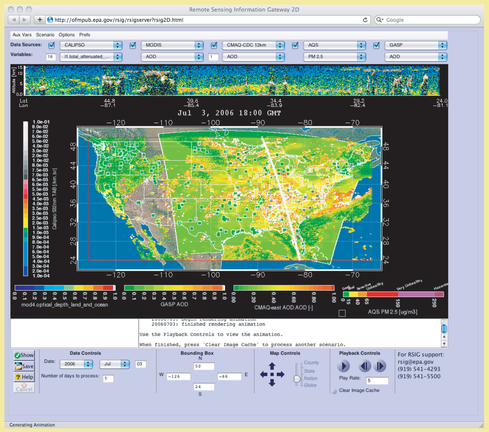
ABSTRACT REFERENCE NUMBER: 1482803
PAPER REFERENCE NUMBER: IN31A-1498

ABSTRACT:

EPA's Remote Sensing Information Gateway (**RSIG, http://www.epa.gov/rsig**) is a widely used free applet and web service for quickly and easily retrieving, visualizing, and saving user-specified subsets of atmospheric data - by variable, geographic domain, and time range.

Petabytes of available data include thousands of variables from a set of NASA and NOAA satellites, aircraft, ground stations, and EPA air-quality models.

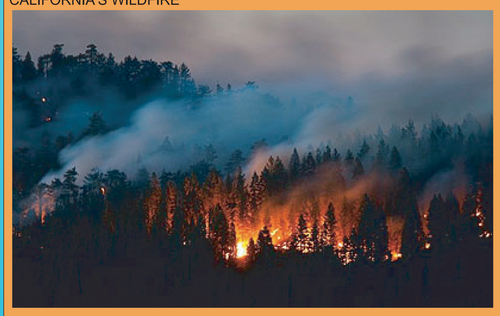
We describe the architecture and technical implementation details of this successful system with an emphasis on achieving convenience, high-performance, data integrity and security.



BACKGROUND/MOTIVATION:

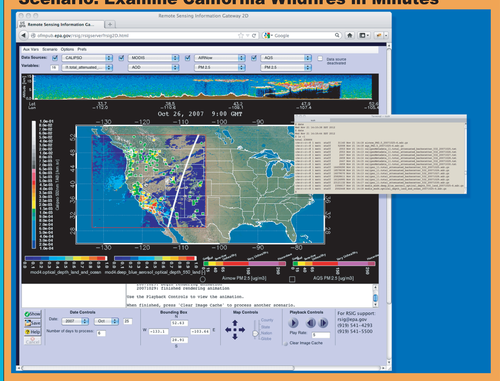
Atmospheric researchers require access to measured data from ground stations, aircraft, and satellites for model evaluation and analysis including exceptional events such as large-scale wildfires.

CALIFORNIA'S WILDFIRE



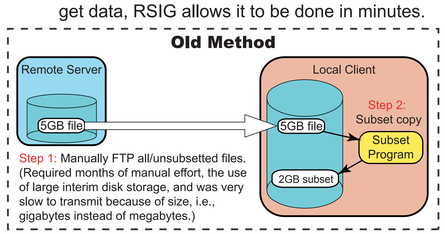
Petabytes of Air Quality (AQ) data are stored at centers around the country, such as NASA's Distributed Active Archive Centers (DAAC) and NOAA's National Climatic Data Center. However, access to these data is often prohibitively difficult, time-intensive, and requires significant staff and computer resources by the consumer.

Scenario: Examine California Wildfires in Minutes



RSIG UNIQUE/KEY FEATURES:

FAST: Where it used to take months of manual effort to get data, RSIG allows it to be done in minutes.



- How long would this have taken using web form+email+ftp?
- How many files would need to be downloaded?
- How many GB of disk space would be needed to store the files?

Data Retrieval: The Old method

To get six days of CALIPSO data for the region of interest:

1. Register with the NASA Goddard DAAC to create an account.
2. Log in to the DAAC web site.
3. After familiarizing yourself with the NASA nomenclature, select the appropriate "project" and "parameters."
4. Wait three days for an email telling you how to download the files.
5. FTP the HDF files.
   In this case, 144 files @ ~0.5GB / file = 72GB and takes 50 hours to download.
6. Process the HDF files into useful, subsetted formats.
   (Only 23 of the 144 files actually contain a track that intersects the continental US.)

Most file-ordering "web−accessible" data sources provide similarly slow, cumbersome, and limited access to the data.

Data Retrieval and Vis: The RSIG Way − Quick & Easy

Visualization

Just five minutes to stream and visualize six days worth of data!
+ CALIPSO LIDAR Backscatter
+ MODIS Aerosol Optical Depth
+ MODIS Cloud Optical Thickness
+ Airnow PM 2.5
+ NESDIS Biomass Burning PM 2.5

Data Download

Saving the data subset to local disk:
Just five minutes to stream & save full−resolution subsetted data and yielded under 250MB in compressed simple−format files.

```
$ date
    Thu Feb 7 15:52:48 EST 2008
  $ date
    Thu Feb 7 15:57:56 EST 2008
  $ ls -asl
        64 -rw-rw-r-- 1 plessel visstaff 31978 Feb 7 15:55 airnow_PM2.5_20071023-6.xdr.gz
         0 -rw-rw-r-- 1 plessel visstaff 0 Feb 7 15:55 aqs_PM2.5_20071023-6.xdr.gz
    473272 -rw-rw-r-- 1 plessel visstaff 242314947 Feb 7 15:55 calipso_TAB532nm_20071023-6.xdr.gz
        16 -rw-rw-r-- 1 plessel visstaff 5319 Feb 7 15:55 goes−bb_PM2.5_20071023-6.xdr.gz
     69000 -rw-rw-r-- 1 plessel visstaff 35298901 Feb 7 15:58 modis2_COT_20071023-6.xdr.gz
      9904 -rw-rw-r-- 1 plessel visstaff 5070301 Feb 7 15:55 modis_AOD_20071023-6.xdr.gz
```

Data Retrieval and Subsetting: The RSIG Method

```
$ uncompress *.gz
$ ls −asl
      224 −rw−rw−r−− 1 plessel visstaff 114367 Feb 7 15:55 airnow_PM2.5_20071023−6.xdr
  1615560 −rw−rw−r−− 1 plessel visstaff 827165389 Feb 7 15:55 calipso_TAB532nm_20071023−6.xdr
      128 −rw−rw−r−− 1 plessel visstaff 61444 Feb 7 15:55 goes−bb_PM2.5_20071023−6.xdr
   292400 −rw−rw−r−− 1 plessel visstaff 149704919 Feb 7 15:58 modis2_COT_20071023−6.xdr
    36232 −rw−rw−r−− 1 plessel visstaff 18548851 Feb 7 15:55 modis_AOD_20071023−6.xdr
```

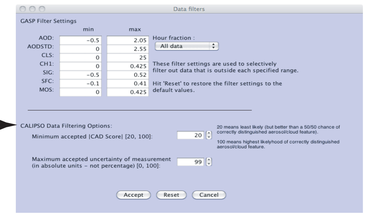
Uncompressed XDR files total just 1GB.

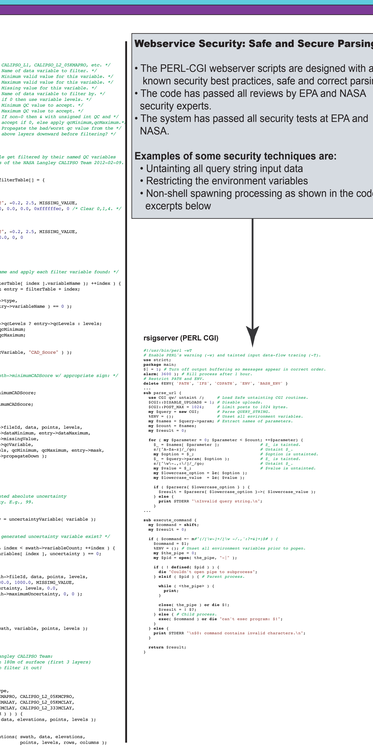
ASCII headers and XDR binary arrays are efficient and easy to read/parse:

```
    $ head −12 airnow_PM2.5_20071023−6.xdr
    AIRNOW 1.0
    subset
    2007−10−23T00:00:00−0000
    # data dimensions: timesteps stations
    144 194
    # Variable names:
    pm25
    # Variable units:
    ug/m3
    # MSB 32−bit integers ids[stations] and
    # IEEE−754 32−bit reals sites[stations][2=<longitude,latitude>] and
    # IEEE−754 32−bit reals data[timesteps][stations]:
```

BARRIERS TO DATA USE INCLUDE:

- Data Discovery: how and where to search for available data useful for particular research needs.
- Learning to use each data provider's website and data-ordering tools, including learning the nomenclature of the various data products.
- Having sufficient local disk storage to hold copies of the ordered, unsubsetted data files. Typically, this will be tens of terabytes.
- Having sufficient time to download a year of typical satellite data, which is composed of hundreds or thousands of large binary data files. Retrieving a year of satellite data could take months.
- Data sleuthing: once the thousands of files have been acquired they must be examined (using binary decoding programs) and deciphered to learn the details.
- Decoding compressed formats (such as various word-size integer encodings).
- **Complex data quality flag interpretation and filtering**





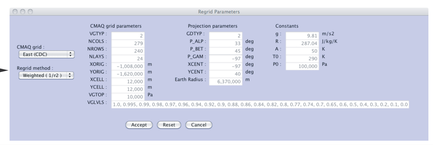
Webservice Security: Safe and Secure Parsing

- The PERL-CGI webserver scripts are designed with all known security best practices, safe and correct parsing.
- The code has passed all reviews by EPA and NASA security experts.
- The system has passed all security tests at EPA and NASA.

Examples of some security techniques are:

- Untainting all query string input data
- Restricting the environment variables
- Non-shell spawning processing as shown in the excerpts below

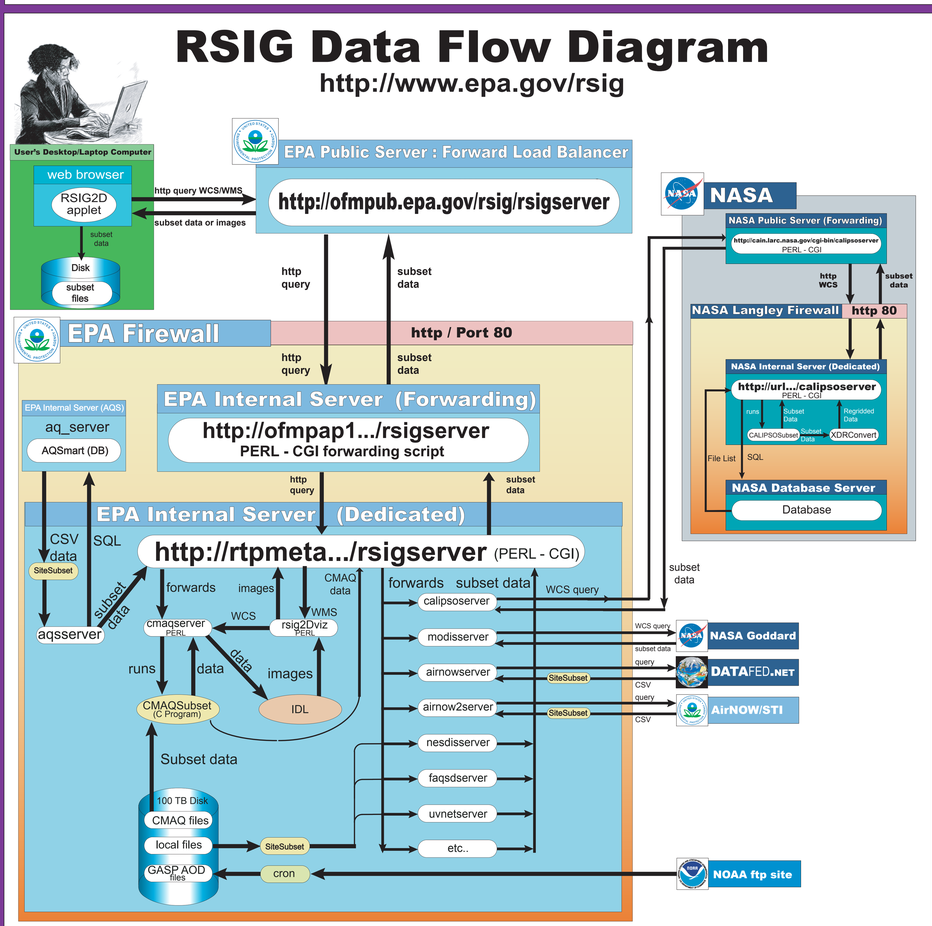
RSIG was created to remove these barriers to use and provide additional value added such as regridded data to modules used by atmospheric researchers at EPA and around the world.



ARCHITECTURE:

Public Components:

- **Applet:** Researchers primarily use the RSIG2D applet whose simple graphical user interface and visualization features enable quick and easy selection, retrieval, visualization, and saving of subsets of air-quality-related data from a variety of sources, including NASA, NOAA, and EPA.
- **Webservice:** The applet uses the free, publicly accessible webservice - rsigserver - to retrieve specified data variables subsetted by longitude-latitude rectangle and date-time range. The rsigserver webservice is based on OGC-WCS for compatibility, data discovery (REQUEST=GetCapabilities) and interoperability with external software applications ("mash-ups").



IMPLEMENTATION COMPONENTS:

Behind the scenes of the public components, there are a chain of webservices invoking data subsetting programs that read the data files needed for a request.

- For large daily-generated datasets, these data-specific webservice applications and subsetters are installed at the data provider site.
- For small or static datasets, the data and processing is stored on a dedicated server inside the firewall at EPA.
- The webserver applications are PERL-CGI scripts that safely parse the query string, issue SQL to a database for the list of data files needed, and invoke the subsetter programs to efficiently read the files - extracting and streaming the subset of data requested back to the EPA server - to be rendered into images for display in the user's web browser or else the data is streamed back to the user's computer and saved to their local disk.
- Subsetter programs are designed for correct data processing, including complex data quality filtering, high performance, and efficiency.

CONCLUSION:

- Development is driven by EPA research needs, as determined by the project's principal investigator and his colleagues.
- Intuitive graphical interface allows users to quickly and easily access and compare selected datasets from massive, remote data repositories.
- Demonstrates the power of collaborative development across Federal Agencies, e.g., NASA and EPA have worked closely together to develop and deploy efficient and secure data server and subsetting codes at the data sources, reducing by orders of magnitude the volume of data streamed over the internet.
- Since project inception in 2005, RSIG has been used by over 100 institutions world-wide and its development continues to evolve with new data and capabilities added every year.

ACKNOWLEDGMENTS:

We gratefully acknowledge the contributions of many who have worked with the RSIG team to provide access to data sets. In particular we would like to acknowledge the support of Danny Mangosing (SSAI), Mark Vaughan, Jason Tackett, Pamela Rinsland, and Chip Trepte at NASA LaRC, and Ed Matsuoka, Steve Kempler, Cid Praderas (Sigma Space), and Greg Ederer (Sigma Space) at NASA GSFC, and Shobha Kondragunta, Pubu Ciren, and Chuanyu Xu at NOAA/NESDIS/STAR.

For more information on RSIG and available data, visit our website: http://www.epa.gov/rsig/

United States Environmental Protection Agency