Geophysical Research Abstracts,
Vol. 11, EGU2009-12428, 2009
EGU General Assembly 2009
© Author(s) 2009

Tools and strategies for instrument monitoring, data mining and data access

Dr R. M. van Hees
SRON, EOS, Utrecht, Netherlands (r.m.van.hees@sron.nl)

The ever growing size of data sets produced by various satellite instruments creates a challenge in data management. Three main tasks were identified: instrument performance monitoring, data mining by users and data deployment. In this presentation, I will discuss the three tasks and our solution.

As a practical example to illustrate the problem and make the discussion less abstract, I will use Sciamachy on-board the ESA satellite Envisat. Since the launch of Envisat, in March 2002, Sciamachy has performed nearly a billion science measurements and performed daily calibrations measurements. The total size of the data set (not including reprocessed data) is over 30 TB, distributed over 150,000 files.

[Instrument Monitoring]
Most instruments produce house-keeping data, which may include time, geo-location, temperature of different parts of the instrument and instrument settings and configuration. In addition, many instruments perform calibration measurements. Instrument performance monitoring requires automated analyzes of critical parameters for events, and the option to off-line inspect the behavior of various parameters in time. We choose to extract the necessary information from the SCIAMACHY data products, and store everything in one file, where we separated house-keeping data from calibration measurements. Due to the large volume and the need to have quick random-access, the Hierarchical Data Format (HDF5) was our obvious choice. The HDF5 format is self describing and designed to organize different types of data in one file. For example, one data set may contain the meta data of the calibration measurements: time, geo-location, instrument settings, quality parameters (temperature of the instrument), while a second large data set contains the actual measurements. The HDF5 high-level packet table API is ideal for tables that only grow (by appending rows), while the HDF5 table API is better suited for tables where rows need to be updated, inserted or replaced. In particular, the packet table API allows very compact storage of compound data sets and very fast read/write access. Details about this implementation and pitfalls will be given in the presentation.

[Data Mining]
The ability to select relevant data is a requirement that all data centers have to offer. The NL-SCIA-DC allows the users to select data using several criteria including: time, geo-location, type of observation and data quality. The result of the query are [i] location and name of relevant data products (files), or [ii] listing of meta data of the relevant measurements, or [iii] listing of the measurements (level 2 or higher). For this application, we need the power of a relational database, the SQL language, and the availability of spatial functions. PostgreSQL, extended with postGIS support turned out to be a good choice. Common queries on tables with millions of rows can be executed within seconds.

[Data Deployment]
The dissemination of scientific data is often cumbersome by the usage of many different formats to store the products. Therefore, time-consuming and inefficient conversions are needed to use data products from different origin. Within the Atmospheric Data Access for the Geospatial User Community (ADAGUC) project we provide selected space borne atmospheric and land data sets in the same data format and consistent internal structure, so that users can easily use and combine data. The common format for storage is HDF5, but the netCDF-4 API is used to create the data sets. The standard for metadata and dataset attributes follow the netCDF Climate and

Forecast conventions, in addition metadata complies to the ISO 19115:2003 INSPIRE profile are added. The advantage of netCDF-4 is that the API is essentially equal to netCDF-3 (with a few extensions), while the data format is HDF5 (recognized by many scientific tools). The added metadata ensures product traceability. Details will be given in the presentation and several posters.