P2.31 A SYSTEM DESIGN FOR STORING, ARCHIVING, AND RETRIEVING HYPERSPECTRAL DATA

Ralph G. Dedecker*, Tom Whittaker, Ray K. Garcia, Robert O. Knuteson
University of Wisconsin-Madison, Space Science and Engineering Center

Hyperspectral data and products derived from instrumentation such as the Atmospheric Infrared Sounder (AIRS), the Cross-track Infrared Sounder (CrIS), Geosynchronous Imaging Fourier Transform Spectrometer (GIFTS) and the Hyperspectral Environmental Suite (HES) will impose storage and data retrieval requirements that far exceed the demands of earlier generation remote sensing instrumentation used for atmospheric science research. A new architecture designed to address projected real time and research needs is undergoing prototype design and development.

1. INTRODUCTION

The large volume of both raw data and products derived from hyperspectral instruments will require large distributed storage devices employing several servers. An operational GIFTS instrument, for example, is expected to deliver approximately 1 terabyte of data per day. The hardware infrastructure must be implemented in such a way so that component augmentation, replacement, and maintenance can be undertaken without necessitating major modifications to user applications. User applications will need tools to simplify locating data files. User data selection facilities for retrieving specific information from storage devices for calibration, analysis, instrument inter-comparison, or reference purposes will also be necessary and standardized data formats and data delivery schemes will be important.

2. CONCEPTUAL DESIGN

Data storage schemes used in support of research efforts at the UW-SSEC using hyperspectral and other remote sensing data have highlighted difficulties associated with having the data distributed over several disjoint workstations. The relatively low volume collected to date using these instruments has already demonstrated the difficulties involved in storing,

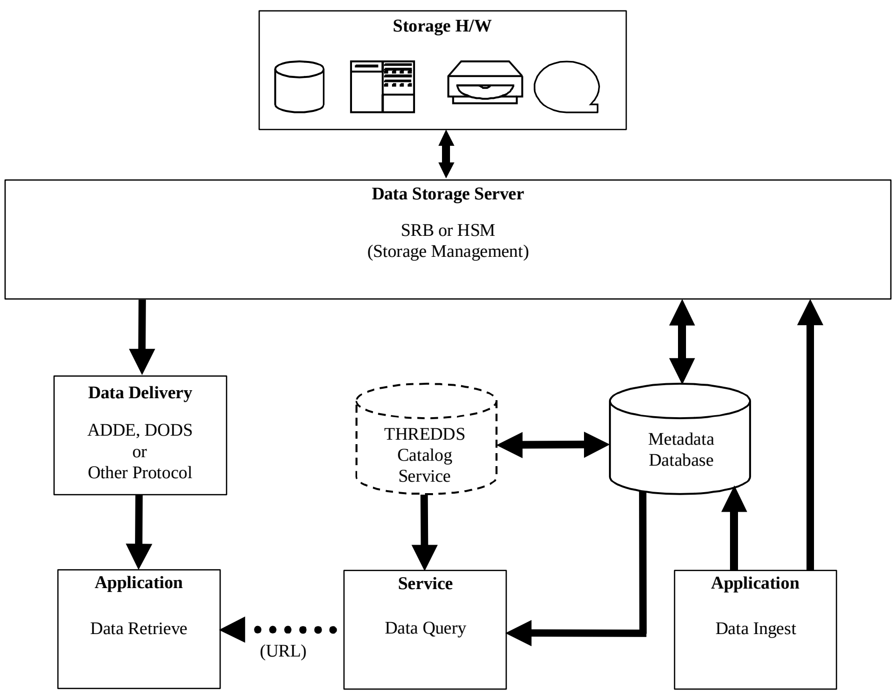

Figure 1: Conceptual Storage System Overview

* *Corresponding author address*: Ralph Dedecker, 1225 W. Dayton St., University of Wisconsin-Madison, Madison, WI 53706; ralph.dedecker@ssec.wisc.edu.

locating, and retrieving data sets. Application software is subject to modifications as data is relocated. The conceptual design described in this paper (see Figure 1) addresses these and other limitations of the current scheme and prepares for the exponential increase in data volume associated with the next generation of instrumentation. The design considerations include:

- Data file storage - distribution, interfacing, maintenance, upgrades, etc.
- Data cataloging
- Database query capabilities
- Data delivery methods

The design goal is to define a hardware infrastructure and software libraries to support the above while adhering to established research application standards where possible. The design goal also includes the ability for vendor-independent component selection so as to harvest future COTS (Commercial Off the Shelf) innovations and flexibility in system integration.

2.1 Data File Storage

The data file storage subsystem (see figure 2) simplifies user application maintenance by decoupling the applications from the data storage devices and the system storage server. Storage hardware includes devices that employ off-line media for long term archives. The subsystem may also use Grid storage schema.

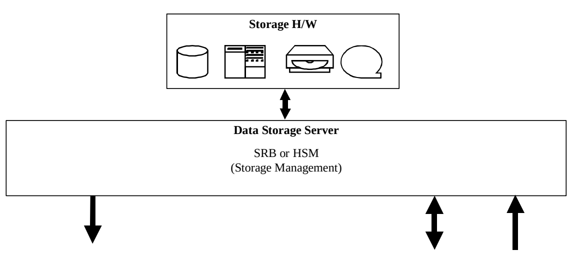

Figure 2: Data file storage subsystem

This subsystem also provides the means for data users (applications and software libraries) to access the data storage infrastructure without dependence on a particular storage device. Via this subsystem, storage devices may be replaced, repaired, and maintained without affecting user software configuration. This subsystem automatically maintains the necessary linkages to data between users of online and archived information.

This conceptual design utilizes storage management support via a commercial HSM (Hierarchical Storage Management) architecture or via open source software schema such as the SRB (Storage Resource Broker) developed at the UCSD.

2.2 DATA CATALOGING

All data files derived from hyperspectral instrumentation and stored by this system will include metadata that will describe these data, any data derived from these measurements, and any data associated with raw measurements or products. Metadata descriptions will also include data quality assessment and links to associated supporting data (such as GOES imagery thumbnails associated with hyperspectral measurements). All data files will be included in a system data catalog that will list data types and selected metadata items. The catalog will be available for browsing. Figure 3 shows the data cataloging subsystem.

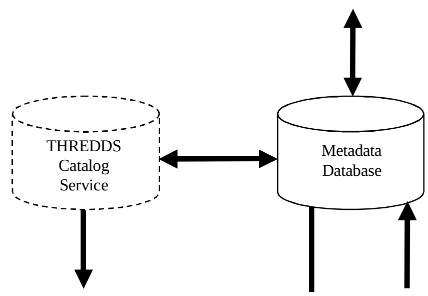

Figure 3: Data cataloguing subsystem

The catalog will be accessible via the THREDDS (Thematic Realtime Environmental Data Distributed Services) standard. THREDDS uses an XML (eXtended Markup Language) based catalog and has UNIDATA & DLESE (Digital Library for Earth System Education) roots.

An XML entry from a THREDDS catalog provides a requesting user program with a URL for accessing the data represented by the catalog entry and a method for visualizing, overlaying, and defining a coordinate system for these.

2.3 Database Query Capabilities

The combined stored data file volume will consist of raw instrument data, derived products, and supporting data. The total volume is expected to increase as case studies are analyzed and results are archived. Total stored data volume is anticipated to be very high and is expected to increase over time. File searches for results and case study data using traditional methodologies would be tedious and cumbersome.

In order to facilitate data selection and searches, the database query subsystem as shown in Figure 4 will support standard SQL (Structured Query Language) operations using the metadata database and data catalogs. User applications will have support libraries available to search the storage system using single or combined metadata items as qualifiers. Returned results will consist of THREDDS XML items directly from a catalog or may consist of simulated THREDDS items for results requiring data combinations.

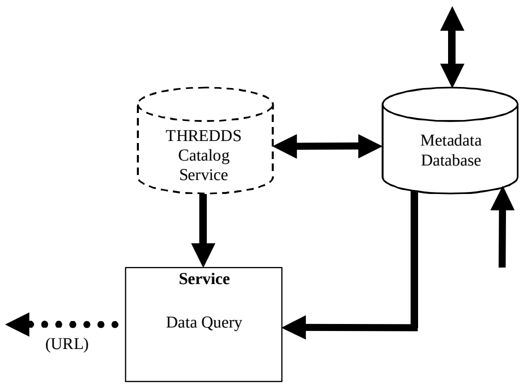

Figure 4: Database query subsystem

2.4 Data Delivery Methods

Requested data will be delivered via the internet or directly to local workstations. Latency in data delivery will vary depending on the type of request and the location of the requested data file. Online data with high bandwidth connections will be delivered in near real-time. For off-line data requests (archived files), requesting applications will be notified of a delay followed by another notification that the requested data is available on line for user transfer via SCP, SFTP, etc.

Figure 5 shows the data delivery subsystem.

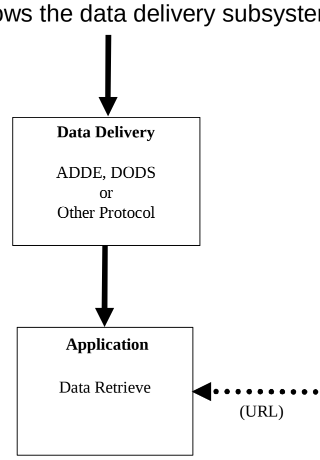

Figure 5: Data delivery subsystem

To facilitate remote delivery and differences in local data formats, standardized protocols and software tools will be employed. Among the delivery mechanisms will be ADDE (Abstract Data Distribution Environment) and OPeNDAP/DODS (Open source Project for a Network Data Access Protocol/Distributed Oceanographic Data System).

ADDE is a client/server mechanism for distributing data. ADDE has UW McIDAS roots. The server provides mechanisms to interpret the data request from the client, retrieve the requested data from disk, arrange the data into the proper format for the client, and send the data back to the application. ADDE is well suited to extract subsets of image, grid, and point data and to provide data transport.

OPeNDAP/DODS is open source software that is very widely used and provides WEB based servers for making local data accessible to remote locations regardless of the local storage format. OPeNDAP/DODS serves data subsets from several formats (HDF, NetCDF, etc.) and thus facilitates interfacing to visualization packages such as Matlab, Ferret, IDV, and IDL.

2.5 Data Submission to the Storage Library

This data storage system is meant for on-line and archive storage of raw measurement data and derived products and information along with associated metadata, indexing, and cataloging information. The storage design does not include features for general data storage or computational scratch storage facilities.

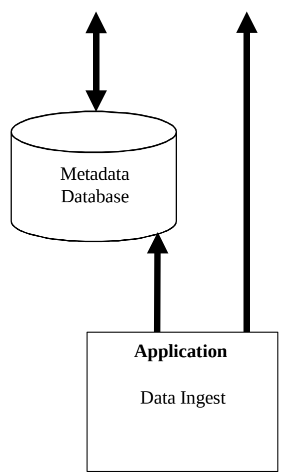

Figure 6: Data submission to library subsystem

In order to meet curator requirements and to maintain hardware and software libraries and stored data linkages, maintenance tool kits will be implemented. Data submission procedures and tools that will protect the integrity and adequate descriptive information of data contributions will be developed. Access to additional tools to assist developer applications software in generating metadata and for data formatting will be provided.

3. IMPLEMENTATION STATUS

This conceptual design is intended to serve as a guideline for efforts to explore operational system design facets through the development and evaluation of prototype systems. The intent is to refine the concepts and to implement an operational data storage system suitable for the next generation of instrumentation. Work at the UW-SSEC has already begun to prototype and apply some of these design concepts using data collected from existing hyperspectral and remote sensing instrumentation.

4. INFORMATION SOURCES

THREDDS:
http://my.unidata.ucar.edu/content/projects/THREDDS/index.html

ADDE:
http://www.ssec.wisc.edu/mug/prog_man/2003/servers.html

OPeNDAP/DODS
http://www.unidata.ucar.edu/packages/dods/index.html

5. ACKNOWLEDGEMENTS

This work was supported by NOAA federal grant NAO7EC0676.