CDAWeb+ Java client: an Experiment in Integrating Disparate Data Systems

Robert M. Candey (Code 612.4)

(Robert.M.Candey@nasa.gov)

Reine A. Chimiak (Code 583)

John F. Cooper (Code 612.4)

David B. Han (Code 586)

Bernard T. Harris (Code 583)

Rita C. Johnson (QSS, Code 612.4)

Colin A. Klipsch (QSS, Code 612.4)

Tamara J. Kovalick (QSS, Code 612.4)

Howard A. Leckner (QSS, Code 612.4)

Michael H. Liu (Raytheon ITSS, Code 612.4)

Robert E. McGuire (Code 612.4)

NASA Goddard Space Flight Center, Greenbelt MD 20771

<http://spdf.gsfc.nasa.gov/>

Abstract

The Space Physics Data Facility (SPDF) at NASA GSFC has developed a strong foundation in space science mission services and data for enhancing the scientific return of space physics research and enabling integration of these services into the emerging NASA Virtual Observatory paradigm.

Our vision is that distributed components of a space physics virtual observatory work together via standard interfaces and metadata agreements to form a globally unified system, comparable to a single super-instrument from a multi-mission ensemble of many data sources. Such a unitary (but not monolithic) view presents geophysical measurements and models across time and space, enabling researchers to easily and seamlessly analyze data from many more sources than possible before.

We are providing a critical set of foundation components, leveraging our data format expertise and our existing and very popular science and orbit data web-based services, such as Coordinated Data Analysis Web [CDAWeb] and Satellite Situation Center Web [SSCweb]. We have added web services for orbit location, data finding across FTP sites and in CDAWeb, data file format translation, and display. These services can now tie together existing data holdings, standardize and simplify their use, and enable much enhanced interoperability and data analysis.

Coordinated Data Analysis Web Java client (CDAWeb+)

- CDAWeb+ is our new Java prototype client built as an experiment in integrating disparate science data services

- Client enables selecting data by combinations of 5 keys (region, mission, instrument type, time span, keyword search)

- CDAWeb+ and its underlying web services and CDAWeb database provide:
  - simultaneous multi-mission, multi-instrument selection
  - comparison of science data via graphics, digital listings, file retrieval and merged/subsetted CDF creation

- CDAWeb comprises >300 datasets of current and past space missions and ground-based facilities (>1M files of science data)

- CDAWeb is very popular:
  - 165k user sessions, 94k plots, 622k FTP requests, 67k ASCII listings, 586 CDF create requests, 2882 file retrieval requests in FY2004

- CDAWeb data and services are available through:
  - FTP file access, including software and documents
  - Web browser (adds data listings, plots, original data files and combined files)
  - Web services (same as above but via SOAP API)
  - CDAWeb+ client for providing additional functionality of underlying CDAWlib IDL library and tying together many other services, including services external to SPDF

- Takes advantage of underlying standards (CDF, ISTP Guidelines, SPASE)

Why build CDAWeb+?

- Integrates disparate services in-house and some external services

- Displays results from detailed queries across many services at once

- Generalizes very popular CDAWeb service to call or point to many services:

    - CDAWeb

    - SSCweb (spacecraft orbit locations)

    - OMNIweb (solar wind, magnetic field and plasma data, energetic proton fluxes, and geomagnetic and solar activity indices)

    - COHOweb/Helioweb (deep space magnetic field, plasma, and spacecraft/planet ephemerides data)

    - ATMOweb (ionospheric and atmospheric data)

    - Modelweb (space physics models)

    - MSQS (Magnetospheric State Query System)

    - FTP Browser (display of subsets of NSSDC's ftp-accessible ASCII datasets)

    - NSSDC Master Catalog (Oracle catalog of information about most spacecraft and instruments)

    - Anonymous FTP data sites (PWGdata, NSSDCftp, etc.)

    - APL Timed GUVI and other PI sites

    - Offline holdings at NSSDC

- Ties together variety of service protocols: FTP, CGI and SOAP web services, Oracle queries, and links

- Makes our existing services more visible and standardizes and simplifies their uses

CDAWeb+ Approach

- Table of manually entered metadata for each dataset or service

- XML version of metadata table, plus other information (such as CDF masters metadata and FTP filename-driven time ranges)

- FTP File Finder concept

  - Using only a few pieces of metadata, we return URLs to files matching a range of time in many FTP and some HTTP data sites

  - Filenaming format based on strptime % strings

  - Example XML (required tags in **bold**)

    <**dataset ID**="ace_mfi_h0_16sec_cdf_cdaweb" nssdc_ID="None"
    serviceprovider_ID="ac_h0_mfi">
    <**access filenaming**="ac_h0_mfi_%Y%m%d_%Q.cdf"
    protocol="ftp"
    subdividedby="%Y" timerange_start="1997-09-02 00:00:12"
    timerange_stop="2004-03-11 23:59:46">
    <**URL**>ftp://cdaweb.gsfc.nasa.gov/pub/istp/ace/mfi_h0</URL>
   </access>   </dataset

- Web Services and SOAP for each service (see below)

- Java and Java 3D for clients

- Java WebStart for easy client install

- IDL on server for data compilation, listing and plotting

- Caveats:

  - No subsetting inventory of given dataset

  - Not attempting to uniformize identification or naming of variables

Web services are key to CDAWeb+ and VOs

- Services Oriented Architecture (vs. client-server)
- Distributed software to software communication, analogous to older technologies such as RPC, DCOM, CORBA, RMI
- No HTML or human interaction required
- Cross-platform and language-independent
- Enables others to develop tools and services leveraging core logic and science data and orbit information
- Everyone can use their own clients/tools

Interoperable web services as basic components of VOs

- Strung together in many combinations
- Form an integrated system much greater than the sum of its parts
- Easily extendable
- Open to other systems and external applications by using standard distributed Application Program Interfaces (APIs)
- Based on XML and Simple Object Access Protocol [SOAP] standards and/or HTTP calling interface
- Tie together existing data holdings
- Standardize and simplify their use
- Enable much enhanced interoperability and data analysis
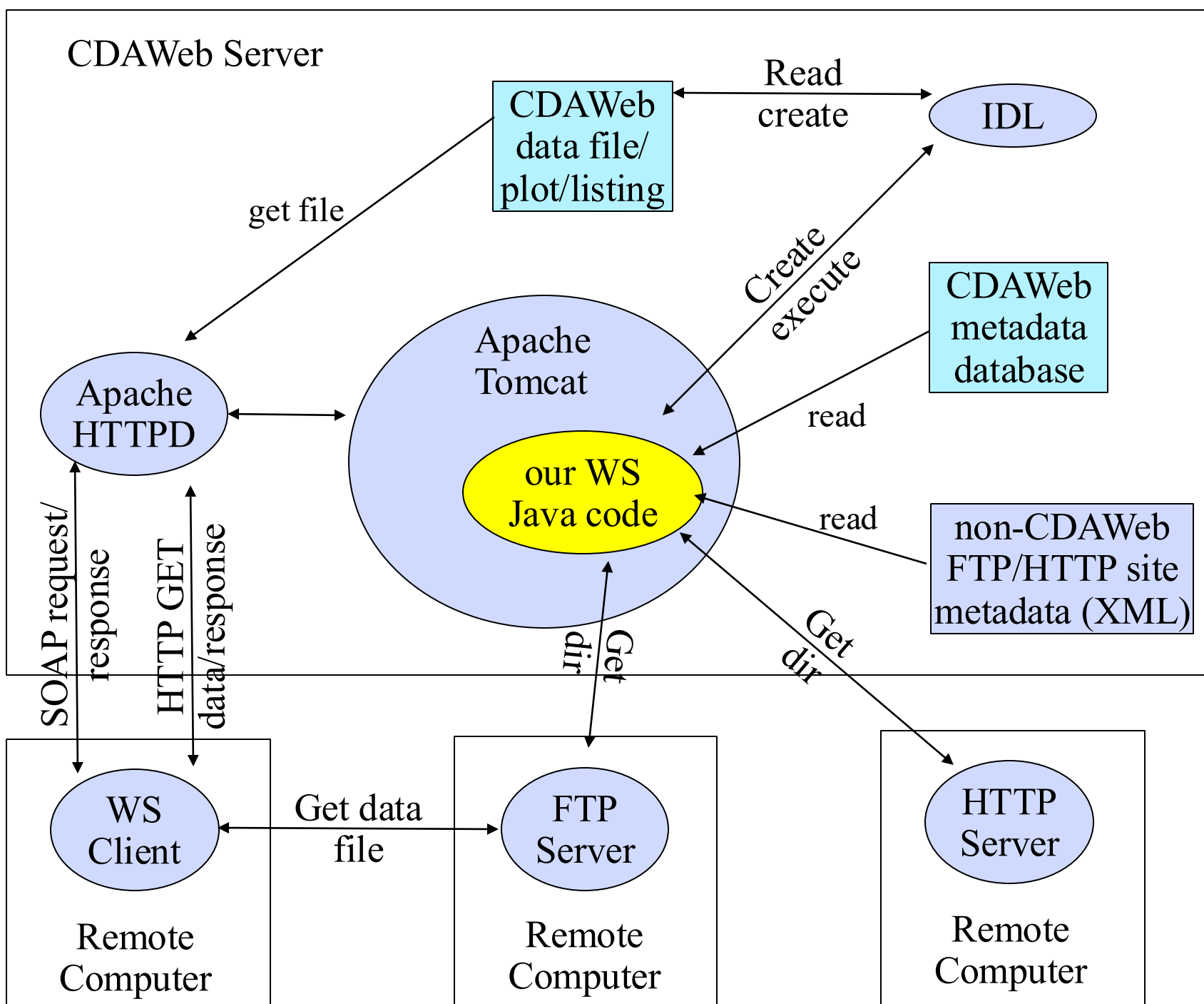
CDAWeb+ Experiment/Questions

Try it soon! -- We need your feedback

- The experiment: How to integrate services of varying functionality?

    - This is an important challenge to VxOs

    - The current range of SPDF services is an interesting testbed

- Possible approaches (for instance, going directly to one spacecraft in the called service and not seeing you can select multiple spacecraft)

    - More comprehensive service calls simpler (subset of features): okay

    - Simpler service calls more powerful (superset of features): perhaps also have pointer to more powerful service main interface to get to additional functionality

    - Partially overlapping set of features between services: how to merge functionality?

- Does CDAWeb+ sufficiently enhance interoperability and data analysis capabilities? For example:

    - Is the dataset-centered paradigm effective?

    - Is it too difficult to use the large lists of results? How else can we shorten it?

    - Are the disclosure triangles effective?

    - Will IT-challenged scientists understand how to use WebStart to start the Java client?

    - Is keeping the navigation window open with other windows popping up helpful?

- How does CDAWeb+ compare to other interfaces (CDAWeb, VSPO, new SPDF Data&Orbits page, etc.)?

- What other search keys should we add? How else to identify data and services? Inventory level or variable level?

CDAWeb+ Concerns

- Metadata population (much is inherently manual and tedious)

  [However, separate metadata provides powerful middle layer for integrated user view]

- Performance and potential load on our servers
- Useful statistics characterizing usage (while preserving privacy)
- Social issues:
  – How to get effective credit (and usage statistics) for services when called by other services
  – How to give credit to other services that you are calling
  – How to assign responsibility to other services being called
- How to call services requiring logins and database queries
- To what extent should we point to external services?
- How to handle incomplete capture of other services and datasets (appear more comprehensive than really are)?
- How do you get a complete list of services in a given domain and maintain it? (distributed domains exponentially harder)

CDAWeb+ Future

- Add parent/child display for grouping related variables (in progress)
- Display of data availability per datasets
- Allow multiple time spans
  – Event lists server (accept/send XML lists, combine lists (and/or), allow annotation, add user-defined or service-defined fields)
  – Bow shock crossings, Magnetopause crossings, etc.
- Add SSC Query (complex multi-spacecraft queries) and OMNIweb extended query (search activity indices) functions
- Allow user more control of plot displays (font, sizes, etc.)
- Add time shifting between datasets to correlate distant spacecraft
- Add plotting to the client (need good Java plot library)
- Sonification (for accessibility and as alternative mode for discovery)
- More pointers to external services (see above concern)??