Search Service Summary

Distributed Infrastructure Design Team

May 31, 2009

Overview

The search service provides functionality for accepting queries from data consumers for data set and data product resources and providing query results. Search services commonly provide a forms-based interface for specifying constraints on any metadata value, a graphical interface for specifying a geographic search extent and full-text searching of descriptive metadata. A search service returns information about matching resources, including a unique identifier that can be used to retrieve the resource. It is common now for searching to be highly interactive, with search options (menu screens) generated based on database contents and limited by previous selections; hits iteratively presented at each search stage; browse representations of a product included with the search result display and results optionally added to a 'shopping cart' for eventual delivery to the user. Some of these capabilities may introduce performance limitations based on the underlying database architecture.

Current practice within PDS

Data set metadata is collected into a high-level catalog database at the Engineering Node. This database requires a substantial overhead to collect, maintain and synchronize. It is presented using either a forms-based search (time, target, mission, instrument, data type) or full-text search on keywords. The search produces a list of data sets links to the appropriate discipline node search sites or to volume browsers. There are a few instances at the discipline nodes where a single data set has its own search system and database. More generally the discipline nodes have gathered all data product metadata into a special database that can be queried with a forms-based, graphical or text search interface. In contrast, the Image Atlas is a generalized search engine that provides parameter driven access to distributed databases that use several different access protocols. One of the main challenges to building the discipline node databases is converting metadata submitted by different instrument teams into a canonical format to allow cross-instrument searches. Most search implementations are using open software including Linux, Java, Tomcat, JSP, AJAX and Google Web Toolkit build on top of MySQL or PostgreSQL databases. One node uses proprietary (Microsoft) components. At one point an effort was made to use the Java-based OODT architecture with distributed query, profile and product servers but only a couple of nodes have deployed services using OODT's complete set of functionality.

Current practice outside PDS

Within EOSDIS each DAAC has its own unique search systems for the data it controls. Metadata for data sets and data products (including browse representations) is submitted to the ECHO clearinghouse where it is loaded in a geospatial database. The database is accessible via several SOAP web service API's, essentially by passing an SQL query to the service. A search client (WIST) is provided but the expectation is that user communities will develop their own specialized search schemes, though progress has been slow. Within the astrophysics community there are several robust search and display applications (e.g. DataScope, Aladin) that access a registry with high-level metadata (discipline, spectral region, resolution) to identify from among thousands of distributed database or data servers then fire off requests to the servers and collate the results. Most server protocols are for a specific data type (table, image, spectra) and search on standard astronomical coordinates and a size field and return a standard XML formatted VOTABLE output. The VOTABLE can be passed on to another analysis program for detailed analysis. The SkyNode service allows specification of a complex SQL query. The space physics and heliophysics communities are moving to a distributed registry and repository architecture where resources are described with standard metadata (SPASE) and can be accessed using SPASE-QL and web services.

Observations

The goal is to have a more uniform product level search across PDS with transparent access to all data while still providing custom search capabilities at the discipline nodes. We see no examples where a standardized search interface has been dictated across a federated system. This alternative would standardize the user experience, but do nothing to provide enhanced access to distributed databases. There seem to be some issues with an exhaustive central database (ECHO) and with a heterogeneous distributed architecture (OODT). Both of these schemes require a substantial learning curve to use the resources. It seems that what is needed is canonical set of data product metadata which will handle the majority of user queries (e.g. observation footprint, spectral characteristics, spatial and temporal resolution) along with a set of simple protocols to deliver the metadata. Regarding search software, the Image Atlas seems to have a good base architecture for providing a general search system for diverse data sets.