Earth Sci Inform (2010) 3:87–94
DOI 10.1007/s12145-010-0050-7

RESEARCH ARTICLE

Mercury: reusable metadata management, data discovery and access system

Ranjeet Devarakonda · Giriprakash Palanisamy · Bruce E. Wilson · James M. Green

Received: 30 September 2009 / Accepted: 14 April 2010 / Published online: 18 May 2010

Abstract Mercury is a federated metadata harvesting, search and retrieval tool based on both open source packages and custom software developed at Oak Ridge National Laboratory (ORNL). It was originally developed for the National Aeronautics and Space Administration (NASA), and the consortium now includes funding from NASA, U.S. Geological Survey (USGS), and U.S. Department of Energy (DOE). Mercury is itself a reusable software application which uses a service-oriented architecture (SOA) approach to capturing and managing metadata in support of twelve Earth science projects. Mercury also supports the reuse of metadata by enabling searches across a range of metadata specification and standards including XML, Z39.50, FGDC, Dublin-Core, Darwin-Core, EML, and ISO-19115. It collects metadata and key data from contributing project servers distributed around the world and builds a centralized index. The Mercury search interfaces allows the users to perform simple, fielded, spatial, temporal and other hierarchical searches across these metadata sources. This centralized repository of metadata with distributed data sources provides extremely fast search results (Table 1) to the user, while allowing data providers to advertise the availability of their data and yet maintain complete control and ownership of that data.

Keywords Metadata · Metadata management · Metadata discovery · Scientific data search · Reusable search engine · ORNL DAAC

Communicated by Thomas Narock

R. Devarakonda (✉) · G. Palanisamy · B. E. Wilson
Environmental Science Division, Oak Ridge National Laboratory,
PO Box 2008, MS6407 Oak Ridge, TN 37831, USA
e-mail: devarakondar@ornl.gov

G. Palanisamy
e-mail: palanisamyg@ornl.gov

B. E. Wilson
e-mail: wilsonbe@ornl.gov

J. M. Green
Information International Associates,
1055 Commerce Park Drive, Suite 110,
Oak Ridge, TN 37831, USA
e-mail: jgreen@iiaweb.com

Introduction

In the recent years, number of scientific data sets created by research projects has increased significantly. However, our current data discovery practices may not be sustainable and reliable as these data sets are spread across thousands of repositories located around the world. Virtual observatories and distributed metadata search and data discovery systems are helping the scientists search those repositories to find and access the required data (Todd et al. 2008). Distributed/virtual metadata systems typically harvest these metadata from various data providers and make it available through a single search system. The Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC) (ORNL 2009), at Oak Ridge National Laboratory developed a distributed metadata harvesting, search and data discovery system called Mercury (Mercury 2008), which was originally developed for searching biogeochemical data that are archived at the ORNL DAAC center. Mercury provides a single portal to information contained in disparate data management systems. It allows investigators/scientists and database managers to distribute their data while maintaining complete control and ownership over it. Mercury is a completely reusable, open-source search system which uses a service-oriented architecture (SOA) to capture and manage biogeochemical and ecological data metadata presently

Table 1 Initial Solr search query time

Search mode	No. of records searched / found	Approx query time (ms)
Full-text search	~70,000/48	90
Fielded search	~70,000/7	122
Fielded search	~9000/4	45
Fielded search	~9000/1382	31

supporting twelve Earth science programs. These programs span a range of land, atmosphere, and ocean ecological communities and have a number of common needs for metadata searches, but they also have a number of needs specific to one or a few projects.

There are other centralized repositories for Earth science data discovery, such as Geospatial One Stop, Global Earth Observation System of Systems etc. As part of data sharing efforts, Mercury provides the harvested metadata to other popular search applications (e.g. Google, Geospatial One Stop, NBII Biobot etc). Some additional benefits to reusing the Mercury software are that it is customizable, inexpensive, and compatible with Internet search engines. Using a consortium approach to development, members share general costs and benefits to reduce development costs and produce a more reusable, portable, robust, feature-rich application. Mercury also provides advance features, such as RSS, Bookmark, External data visualization tools integration, faceted filtering options, etc. In this paper we discuss Mercury's harvesting models, indexing techniques, and various search services that are available through the Mercury system.

Methods and techniques

Mercury supports widely used metadata standards such as Federal Geographic Data Committee (FGDC) (FGDC 2009), Dublin-Core, Darwin-Core, Ecological Metadata Language (EML), Directory Interchange Format (DIF), and ISO-19115, ISO-19139, Open Archives Initiatives-Protocol for Metadata Harvesting (OAI-PMH) (OAI 2009) and protocols and specifications such as XML and Z39.50. It is based on open source and Service Oriented Architecture and provides multiple search services.

Mercury's architecture has three major reusable components in (Fig. 1): a harvester engine, an indexing system, and a user interface component. The harvester engine is responsible for harvesting metadata records from various distributed servers around the world. The harvester software has been packaged in such a way that all Mercury projects use the same harvester scripts but each project will be driven by a set of project specific configuration files. The harvested files are structured metadata records that are indexed against the search library API consistently, so that it can render various search capabilities such as simple, fielded, spatial and temporal.

Harvester

Mercury's harvester operates in two different models, 1) virtual internet database and 2) virtual aggregate database.

Virtual internet database

The virtual internet database model organizes a new collection of data from informal systems spread across the internet, Typically the data providers or the principal investigators create the metadata for their data sets using their own applications or ORNL's Metadata editor and place these metadata in a publically accessible place such as a web directory or FTP directory. Mercury then harvests these metadata from several different contributing agencies and builds a centralized index and makes it available via the Mercury search interface (Fig. 2). Frequent, automated harvesting and complete rebuilding of the index keep the aggregate database up-to-date.

Virtual aggregate database

In the virtual aggregate database model, Mercury harvests information from existing formal disparate database management systems (DBMS). In this model, the metadata exists in remote databases, custom export programs can be easily written to extract the metadata from these DBMS and the metadata are saved in xml files. Mercury then harvests the extracted metadata files and builds a centralized index for metadata searching (Fig. 2). Some Mercury instances are using both these models to harvest the metadata. Mercury development team is currently working on enabling a metadata harvesting service using the OAI-PMH.

Searching

The current version of Mercury (V3) was recently redesigned to be based on open source, rather than commercial, technologies for the indexing and searching components. Lucene (Lucene 2009) and Solr (SOLR 2009a) search libraries, which are both part of the Apache project, form the basis for this part of Mercury. Lucene is rated as the top 10 open source projects (Census 2009) and one of the top 5 Apache projects (Apache 2009). Solr is built on top of Lucene, which is a proven technology over 8 years old with a huge user base (Lucene-java 2009) (this is only a small part). Solr extends the functionality by providing for numeric data types, dynamic fields, unique keys, and faceted searching. For Mercury, Solr provides the capability

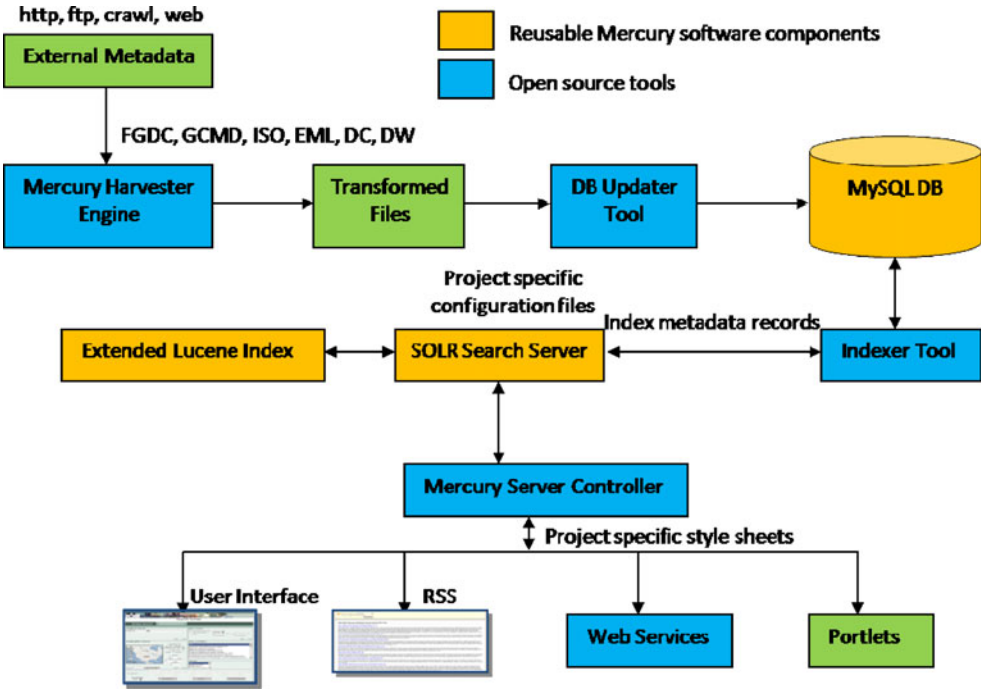

Fig. 1 Mercury system architecture

for specific geo-temporal coordinates, enabling bounding box types of spatial metadata queries. Solr can also treat special information, such as advanced search fields and field importance weighting, as opposed to using the default Lucene rankings and search mechanisms.

Another popular feature of Solr is its built-in caching capability (Solr 2009b). Unlike other caches, Solr caches don't expire after a certain period of time, rather, caches are preserved until the Solr index is valid. New searches can auto-warm previous searches while the current one is still serving external requests. So, if Solr encounters the same query more than once, it returns the results in 0 ms. This feature gives Mercury a performance boost while searching for thousand of records. Table below is a sample of initial Solr query response times:

Results and discussion

The typical Mercury user interface provides three different search capabilities. 1) Simple search, 2) advanced search and 3) Hierarchical (web brows tree) search. In the simple search option, users can perform a full text search. In the advanced search option, users can search by specifying keywords, time period, spatial extend and the data provider information. In the web browse tree, users can drill down through all of the metadata records based on a project-specified hierarchy of fields and filters. Figure 3 is a snapshot of the Mercury advanced search interface used in ORNL DAAC (ORNL DAAC Mercury 2009), and Fig. 4 shows an example of the web browse tree interface.

Once the users enter their search criteria and perform the search, the results summary page displays the total number of records found for the search and options for filtering the search results using logical groupings (by data providers, parameter, sensor, topic, project etc.). The summary page also allows the users to sort the results based on the search relevancy, period of record, source and project. The page shows push buttons in the top right to create an RSS feed, a bookmark or an email for these results.

The bottom of the summary page shows the results, snippets of the records that match the search/browse criteria, and a link to the full metadata and a link to access the associated data. The stars shown at the bottom of each record indicate the relative relevance of the matched criteria. The snippet includes the title and study date range, source provenance and excerpts from the abstract (Fig. 5).

When the user clicks the "View Full Metadata" link found on the summary page, the Mercury metadata report's page will be displayed. This page offers two styles to display a full metadata record. The Mercury by default offers a classic, well organized redux style at the full records page. Additionally, it offers what it is known as the FGDC style, which should be very familiar to those who use the ESRI tools or that have used the previous version of Mercury. It is plain text divided in 6 sections, with the underlying hierarchy preserved as indentation.

Users can create a bookmark, email their custom search results or subscribe to an RSS feed. RSS and bookmarks enable refreshing the query results periodically without the need to recreate the original query. For example, if the user searches for "soil temperature" in the (Long Term Ecological

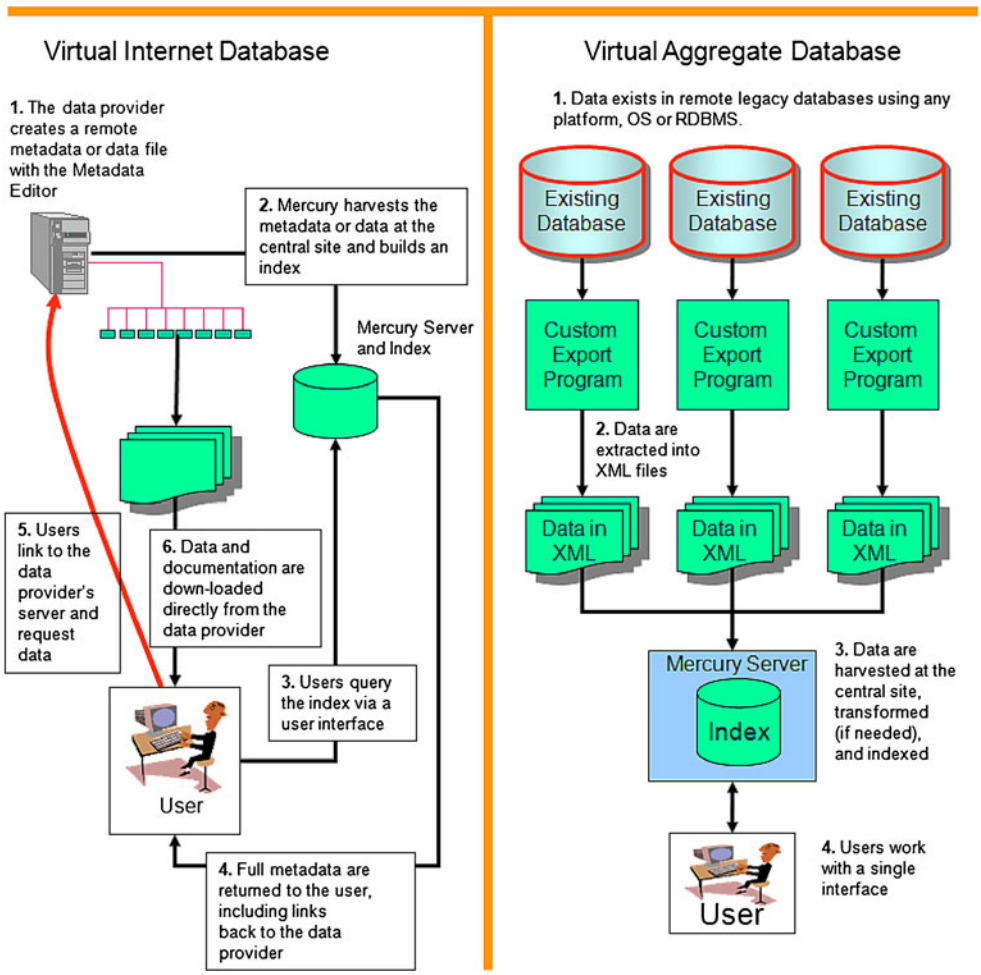

Fig. 2 Mercury metadata harvester's architecture

Resource) LTER data source, Mercury will return references to the 78 LTER metadata records which contain "soil temperature" in the metadata record. The user can then select the RSS button (Fig. 6) on the result summary page to get the RSS URL for this specific search criteria (i.e., full text: soil temperature and data source: lter), an example URL would be similar to the following: http://mercury.ornl.gov/ornldaac/send/processRss?term1=soil+temperature&term1attribute=

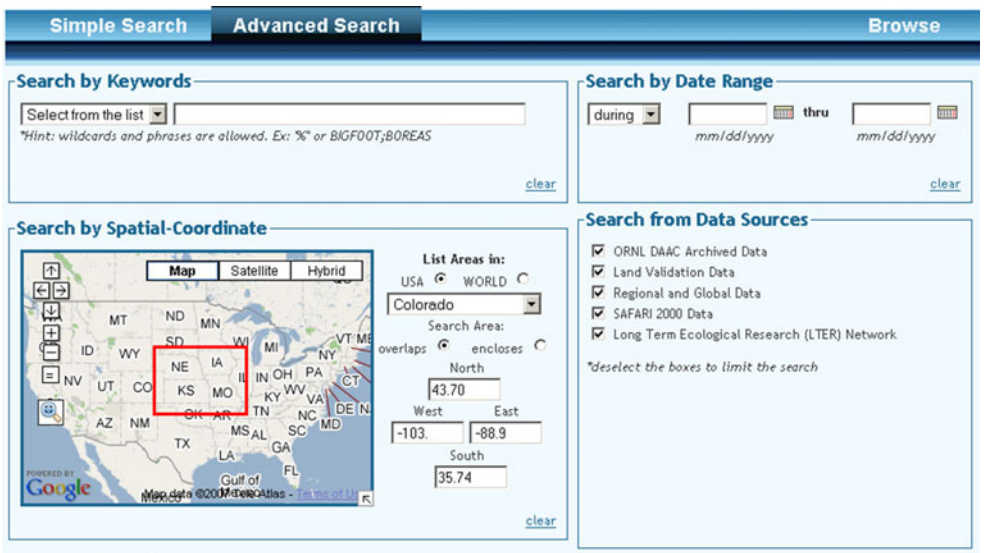

Fig. 3 A snapshot of the ORNL-DAAC advance search interface

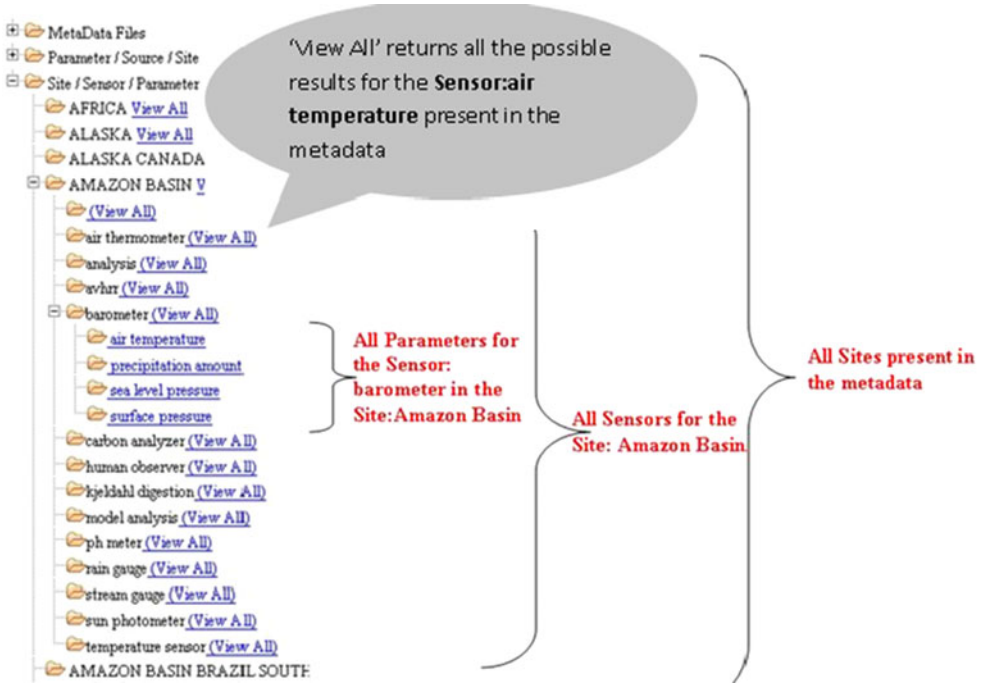

Fig. 4 A snapshot of the ORNL-DAAC browse tree search architecture

Fig. 5 A typical look at the query results page

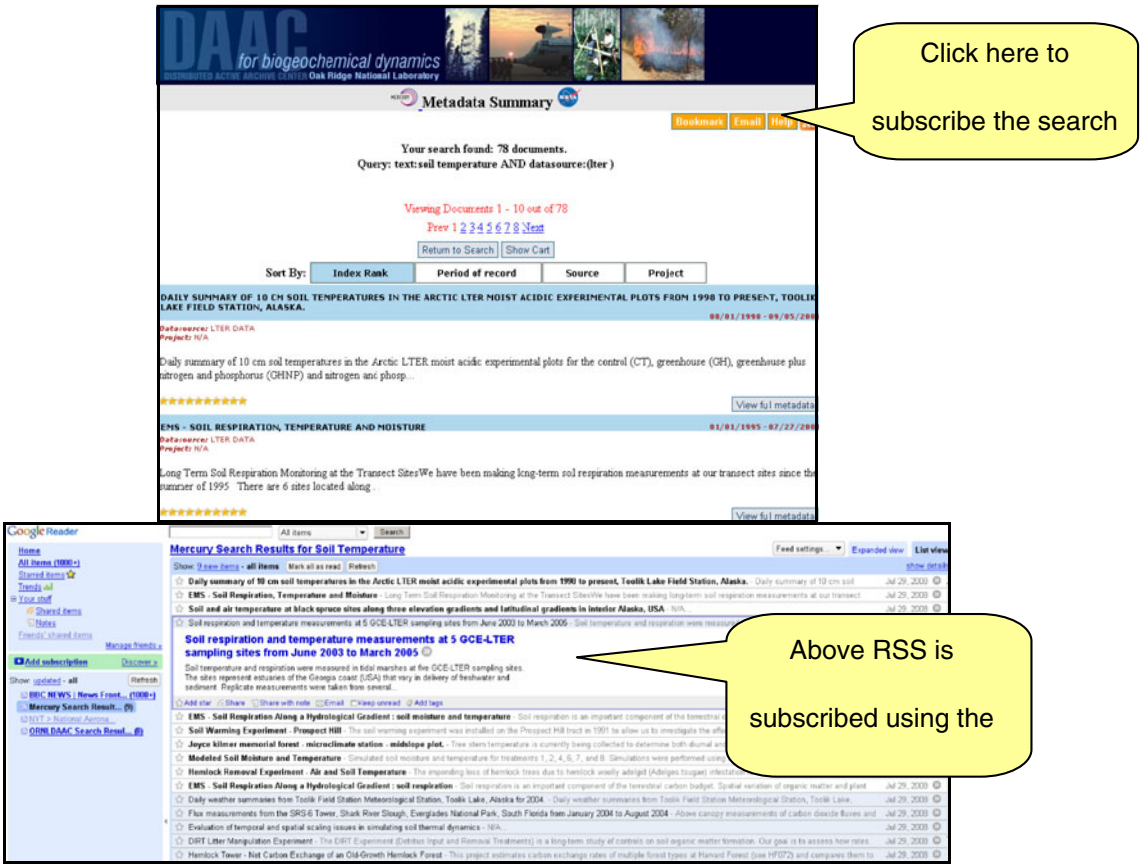

Fig. 6 Creating RSS feed from a search result

text&op1=&term6attribute=datasource&op6=+OR+&term6=lter&pageSize=10&start=0&sortattribute=default&sortattribute=default. Users can then use this RSS URL in any of the many RSS readers (e.g., Google Reader, iGoogle, MyYahoo etc.,) that are available online for subscribing to search results. Whenever the RSS reader refreshes the feed, Mercury will perform a new search and provide the latest search results, and the newly added records will be displayed at the top. The user can obtain the full metadata report by selecting the link found in the RSS feed.

Mercury also provides the harvested metadata to other applications (e.g., Google, NASA Global Change Master

Fig. 7 Snapshot from mercury—discovery, access, and delivery of data for IPY (DADDI) search

Directory, and NBII Biobot).The National Biological Information Infrastructure (NBII 2009a) (NBII) Metadata Clearinghouse (NBII 2009b) consumes the search results as portlets in their NBII portal web application, which is another way of displaying the customized search results in external web pages. One of the partners in the NBII Clearinghouse, Global Forestry Information Services (GFIS) (GFIS 2009) is harvesting all the forest related metadata records as RSS service and exposing those records through their search system.

Cross-platform external tools reuse

Mercury allows seamless integration with various external software tools for data visualization, data ordering etc. This is an important and useful approach to a software reuse. Mercury not only supports classic data-integration paradigm, i.e., having a common data format and/or using a segment of source code, but it also support multi-platform computing. An example to clarify what this means: Mercury reuses remotely located Socioeconomic Data and Application Center (SEDAC)'s Web Map client as a visualization tool for the user requested data (Fig. 7). Reusing components like these not only provides additional data usability options, but also tremendously reduces duplication of programming efforts, development costs.

Ongoing and future development

Continuous efforts are being made at ORNL DAAC to the Mercury codebase to improve the quality of the tools and to expand the scope to add new services. Currently, work is being done to support various Web services activities such as Online Web Thesaurus service, Gazetteer web services, and Open Geospatial Consortium (OGC) Catalog service. Integration of frameworks such as Open-source Project for a Network Data Access Protocol (OPeNDAP) (OPenDAP 2009), OpenSearch (OpenSearch 2009) and Search/Retrieval via URL (SRU), and Web service harvesting and provider web services are also included in the Future development plan.

Conclusions

Mercury indexes and searches more than 50,000 metadata records through its various project specific user interfaces. Mercury supports various metadata standards including XML, Z39.50, FGDC, Dublin-Core, Darwin-Core, EML, and ISO-19115. The Mercury system has a completely reusable open source Service Oriented Architecture and provides multiple search services including; user interface search tools, RSS services for search results, bookmark search results, portlets supports.

Acknowledgements Oak Ridge National Laboratory is managed by the UT-Battelle, LLC, for the U.S. Department of Energy under contract DE-AC05-00OR22725

Mercury was presented in many conferences including 2007 (Palanisamy et al. 2007) and 2008 (Devarakonda et al. 2008) American Geophysical Union (AGU).

The Mercury consortium is funded by NASA, USGS, and DOE for a consortium of projects, including ORNL DAAC, NBII, DADDI, LTER, The Large Scale Biosphere-Atmosphere Experiment in Amazonia (LBA), Better Air Quality for North America (NARSTO), Carbon Dioxide Information Analysis center (CDIAC), IABIN Invasives Information Network (I3N) and Inter-American Institute for Global Change Research (IAI).

References

Apache Stats (2009) Daily stats. The apache XML project. Available at: http://people.apache.org/~vgritsenko/stats/daily.html. Accessed September 2009

Census Reports (2009) Census summary report. The open source census. Available at: http://www.osscensus.org/summary-report-public.php. Accessed September 2009

Devarakonda R, Palanisamy G, Green J, Wilson BE (2008) Mercury: an example of effective software reuse for metadata management, data discovery and access, AGU, 89(53), Fall Meet. Suppl., IN11A-1019

FGDC (2009) Federal geographic data committee (FGDC). Available at: http://www.fgdc.gov/.Accessed September 2009

GFIS (2009) Global Forest Information Service home page. International Union of Forest Research Organizations. Available at: http://www.gfis.net/. Accessed September 2009

Lucene (2009) Apache lucene—overview. Lucene. Available at: http://lucene.apache.org/java/docs/index.html. Accessed September 2009

Lucene-java Wiki (2009)—User base, apache, available at: http://wiki.apache.org/lucene-java/PoweredBy. Accessed September 2009

Mercury (2008) Distributed metadata management, data discovery and access system. Oak Ridge national laboratory. Available at: http://mercury.ornl.gov. Accessed August 2009

NBII (2009) National biological information infrastructure home page. Biological informatics office of U.S geological survey. Available at: http://www.nbii.gov/. Accessed September 2009

NBII CH (2009) National biological information infrastructure metadata clearinghouse. U.S geological survey, NBII, ORNL, Mercury. Available at: http://mercury.ornl.gov/nbii. Accessed September 2009

OAI (2009) Open archives initiative—protocol for metadata harvesting. Available at: http://www.openarchives.org/pmh/. Accessed September 2009

OPenDAP (2009) Open-source project for a network data access protocol home page. Available at: http://opendap.org/index.html Accessed September 2009

OpenSearch (2009) Open search organization home page. Available at: http://www.opensearch.org/Home. Accessed September 2009

ORNL DAAC (2009) Distributed active archive center for biogeochemical dynamics home page. Oak Ridge national

laboratory. Available at: http://daac.ornl.gov/. Accessed September 2009

ORNL DAAC Mercury CH (2009) Oak Ridge national laboratory distributed metadata management, data discovery and access system mercury. Oak Ridge, ORNL, Mercury. Available at: http://mercury.ornl.gov/ornldaac/. Accessed September 2009

Palanisamy G, Wilson BE, Devarakonda R, Green J (2007) Mercury search system, mercury: distributed metadata management, data discovery and access system American Geophysical Union, Fall Meeting 2007

SOLR (2009) Apache Solr main page. Lucene. Available at: http://lucene.apache.org/solr/. Accessed September 2009

Solr Wiki (2009)- Solrcahing. Avialable at: http://wiki.apache.org/solr/SolrCaching. Accessed Feb 2010

Todd King, Narock, T, Walker, R., 2008. A brave new (virtual) world: distributed searches, relevance scoring and facets 1:29-34. doi: 10.1007/s12145-008-0002-7