Lightweight Advertising and Scalable Discovery of Services, Datasets, and Events
Using Feedcasts and Social Tagging

Brian Wilson, Gerald Manipon
Jet Propulsion Laboratory, California Institute of Technology

Rahul Ramachandran
University of Alabama Huntsville

Introduction

NASA's Earth Observing System (EOS) is the world's most ambitious facility for studying global climate change. The mandate now is to combine measurements from the instruments on the "A-Train" and numerous smaller Earth probes to enable large-scale studies of climate change over periods of years to decades. However, moving from predominantly single-instrument studies to a multi-sensor, measurement-based model for long-duration analysis of important climate variables presents serious challenges for large-scale data mining and data fusion. To efficiently compute *climate data records* and study "events" in the climate, one must locate the datasets from multiple sensors at various data centers, query for data granules in space & time, and access terabytes of data. Much of this discovery, space/time query, data access, and analysis work can be performed using Web Services provided by the various data centers (NASA DAAC's) and the EOS Clearinghouse (ECHO). Unfortunately, many of these services are not well known or are poorly described, and therefore are under-utilized.

Discovery and use of Web Services for querying, accessing, and processing Earth Science datasets is hampered by the lack of an ***open and web-scalable*** services and data registry that provides rich and complete search for all available services, datasets, interesting geophysical events, and data granules relevant to studying those events. Existing registries, like NASA's Global Change Master Directory (GCMD) or the Earth Observing System Clearinghouse (ECHO) provide limited search capabilities (often only text keyword) for data collections and services, with no support for events and linked data granules. The Group on Earth Observations System of Systems (GEOSS) Registry provides text keyword search for registered standards and services (e.g. data query/access services) and spans international and institutional boundaries. Such centralized registries are useful for basic metadata search but due to their inherent paradigm they suffer from various failings: they often have cumbersome interfaces for registering services or datasets, thereby presenting a barrier to adoption; each use their own metadata standard, which may not be easily machine-readable or interoperable; they don't evolve their metadata fields (registered information) rapidly to suit changing user needs; their search results are often a long catalog without relevance ranking to help the user; and they compete with each other for adoption, thereby fragmenting the user base. To find all available web services, the user must search all three registries (do meta-search), and even then the resulting set will omit many unregistered services and datasets.

Our Approach

Under a NASA ACCESS grant, we are developing a lightweight service advertisement mechanism that will improve web discovery of provider's Earth Science services, while also being backward-compatible by providing search over existing metadata repositories. By combining lightweight service casting with meta-search, we will provide a novel and a complete solution:

1. Provide a lightweight, easily-authored mechanism for ES service producers to publish and promote their services;
2. Enable service consumers to easily find & invoke services, both those published as service casts and those registered in the GCMD, ECHO, and GEOSS metadata repositories.
3. Provide a rich search interface as a browser plug-in that has modern, web 2.0 capabilities: search term suggestions for the user, term synonyms and broadening using semantics, search by user tags, and integrated social tagging [*Smith, 2008*] so users themselves can enhance the categorization of services, rank them by usefulness or performance, tag them for use in collaborative service chains, etc.

The objective is to provide a **one-stop search box** where users can search by keyword and service taxonomy. For example, a user might want to find water vapor data from the Atmospheric Infrared Sounder (AIRS) on the

Aqua satellite. Searching for “airs water vapor” will return relevance-ranked results for AIRS **metadata** in GCMD & ECHO **and services** that provide query & access to AIRS datasets, like the MIRADOR search services maintained at the Goddard DAAC. Or by searching for “clustering”, one can browse data mining services that can perform clustering, like some available in the ADaM data mining suite at UAH. (In all cases, the results will cover all advertised (cast) services and all registered services.) The user can then tag a service by its features and rank its performance. As more users tag, the collective intelligence will grow so that highly-tagged or highly-ranked services are featured in the relevance ranking, thereby enriching service discovery.

Syndication feeds have already been applied in the Earth science domain to advertise the existence of new data granules are they are produced from space-based sensors. Datacasting was developed under a prior NASA ACCESS grant by JPL’s Andy Bingham and Tim Stough (Co-I) [*Bingham et al, 2008*; see http://datacasting.jpl.nasa.gov]. We have already begun promulgating the lightweight “service casting” approach in which services are advertised on the web in Atom syndication feeds, which are searchable at Google (the Feed API) and can also be aggregated to provide smart, faceted search (semantics beyond text keywords). The serv-cast (Atom XML) v1 standard is currently being vetted by ESIP Federation members and a catalog of published “servcasts” is already being accumulated. Servcasts can be used to search by service taxonomy, look up the service interface (e.g. WSDL/WADL), machine auto-invoke the service, or click through to service documentation for humans. If a provider’s services change or expand, the advertisement can be modified and re-cast. The service provider is in control, and can add metadata fields, or even information intended for its own use, to the extensible servcasts at any time.

Similarly, metadata for data collections can be published in “dataset casts”, and interesting geophysical events (e.g. hurricane tracks) or studies of periodic structures (e.g. El Nino), with links to related datasets and services, can be advertised in “event” (or interest) casts. Under this work, we are defining the dataset and event cast standards, which leverage existing ISO and XML metadata standards. We plan to populate a body of discoverable service and event casts by harvesting collection and services metadata stored in the GCMD, ECHO, and GEOSS registries. The three casting standards—service, data, and event—are collectively referred to below as “infocasts” (short for metadata broadcast using feeds).

In addition to defining and publishing the infocasting standards, we are also creating a “social” browser application for discovery of web services, data collections, and interesting geophysical events that, by aggregating casts and utilizing Web 2.0 technologies, provides text keyword search, faceted (semantic) search, interesting new services & events of the day, recommendations for useful services or big events, and search by user tags. Users will be able to tag services and events/datasets using a pre-defined hierarchical taxonomy or their own categories, publish new events, link datasets to events, and rank services (one to five “stars”) by suitability for purpose, performance, availability, usefulness of outputs, etc. The semantic search will be implemented by extending Noesis technologies developed by Ramachandran [*Ramachandran et al.,2006; Movva et al.*], and by leveraging built-in features of the Drupal Content Management System (CMS), including the social tagging module and the recommendation engine.

Lightweight Service Casting

Each entry in a service cast advertises a bundle of web services (e.g. SOAP or REST services), described by a machine-readable interface description (e.g. a WSDL or WADL document), callable at a service endpoint URL, and accompanied by links to human-readable documentation. All of this information is machine-readable so that a service can be discovered and auto-invoked by an automated script that understands the service protocol (e.g. OGC/WMS, OpenDAP, opensearch, etc.).

Figure 1 below shows a “servcast” entry for a SOAP service that provides space/time granule query for several datasets. Each Atom feed entry contains the usual metadata fields (XML elements) such as title, id, update time, and text summary. The pointers to the interface description, server endpoint, and documentation are structured (machine-readable) ‘links’ with (purposed) types (rel=) scast:interfaceDescription, scast:serviceEndpoint, and scast:serviceDocumentation respectively. In Atom one can also use links with 'rel=enclosure' for enclosures or the ‘content’ XML tag for in-line content. The 'content' tag can contain HTML or any XML information from additional namespaces. In Figure 1, it is used to show an example of how to call the service (in HTML-formatted text). The complete list of servcast metadata fields is summarized in Table 1 below.

```
<entry>
```

```
  <title>GeoRegionQuery</title>
  <id>uri:http://scifo.jpl.nasa.gov/sciflo/v1/services/GeoRegionQuery</id>
  <updated>2008-03-13T01:32:02Z</updated>
  <summary>Space/time query and granule URL lookup services for multiple EOS
L2/L2 datasets:  AIRS, MODIS, MISR, GPS, and AERONET (ground network).
  </summary>
  <scast:serviceSemantics>Simple</scast:serviceSemantics>
  <scast:serviceProtocol>SOAP</scast:serviceProtocol>
  <category schema="scast" term="sciflo data query space time" />
  <link type="application/wsdl+xml" title="Service interface description”
href="http://sciflo.jpl.nasa.gov/services/soap/2006v1/EOSServices.wsdl" />
  <link rel="scast:interfaceDescription" type="application/wsdl+xml"
     title="Service interface description"
     href=" http://sciflo.jpl.nasa.gov/soap/2006v1/EOSServices.wsdl" />
  <link rel="scast:serviceEndpoint" type="application/soap+xml"
     title="Server endpoint"
     href="http://df3.jpl.nasa.gov/sciflo/services/soap" />
  <link rel="scast:serviceDocumentation" type="text/html" xml:lang="en-us"
     title="Service documentation"
     href="https://sciflo.jpl.nasa.gov/SciFloWiki/SpaceTimeQuery" />
  <content type="xhtml">
   <div xmlns="http://www.w3.org/1999/xhtml">
    <p><b>Example Call</b></p>
    <code>GeoRegionQuery(dataSetId='AIRS', level='L2', version=None,
       startTime='2006/01/01T00:00:00', endTime='2006/02/01T00:00:00',
       latMin=-90., latMax=90., lonMin=-180., lonMax=180.,
       responseGroups='Large')
    </code>
   </div>
  </content>
 </entry>
```

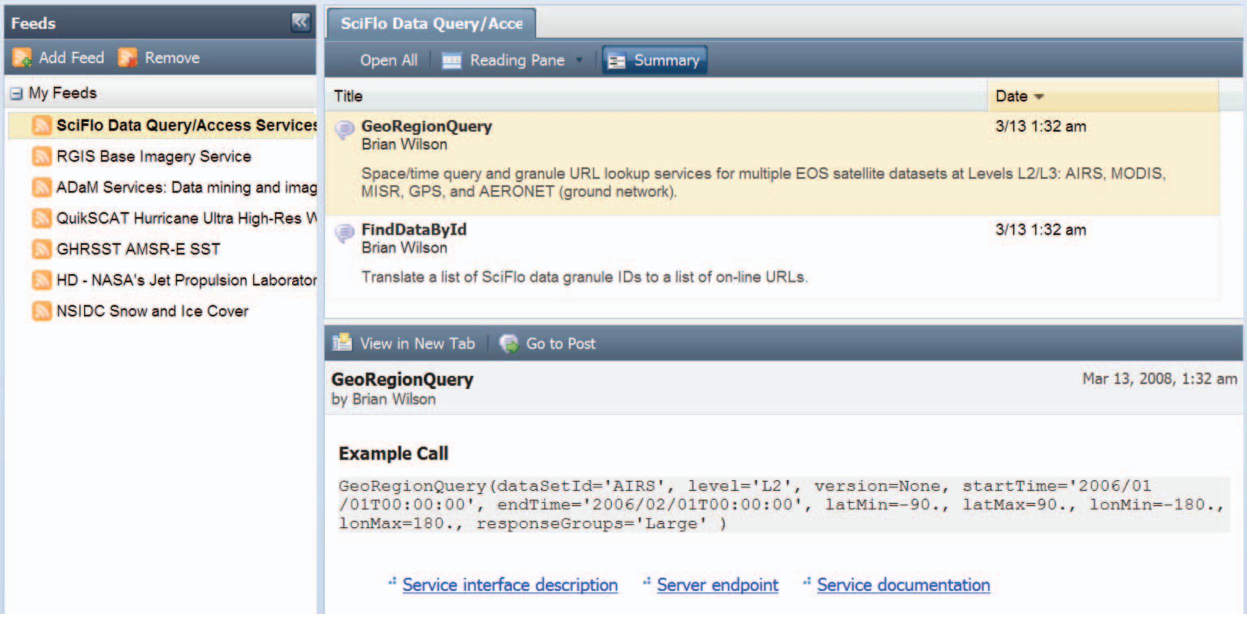
Figure 1. Example entry in a servcast, illustrating tags in the ‘scast’ namespace and the three typed links to machine-readable interface, service endpoint, and human-readable documentation.

Table 1. Metadata Fields in each Service Cast Entry.

Metadata Field	Purpose	Example Values
Entry title, URI, last update time, and text summary	Searchable metadata	Text strings and ISO date & time, e.g. 2008-03-13T01:32:02Z
Service semantics (values from taxonomy)	Services with known semantics can be machine-scripted	OGC.WXS, OpenDAP, Opensearch, AdHoc, Simple, & Human
Service protocol or invocation syntax (values from taxonomy)	Messaging protocol (“plumbing”) used to invoke the service	SOAP, REST, HTTP, AJAX, & JavaWebStart
Link to interface description	Enable machine-parsing of service interface	URL pointing to WSDL or WADL doc.
Link to service endpoint	Enable script to auto-invoke the service	URL pointing to server
Link to human-readable documentation	Default link in feed reader	HTML docs.
Embedded ‘content’ tag (HTML or XML)	Provide further in-line explanation	Example of a call to the service
Category information	Service taxonomy terms provided by scast author	E.g. “data mining clustering”
User tags	Additional taxonomy terms and ad hoc tags provided by users of services	E.g. “AIRS L2 variable water vapor”

Further information about service casting is available at http://sciflo.nasa.gov/scast. A more extended example of a servcast for several of SciFlo's [*Wilson et al., 2005*]data query/access services (e.g. GeoRegionQuery and FindDataById) is viewable at http://sciflo.jpl.nasa.gov/scast/Sciflo_df3.scast.atom. You can load the servcast into virtually any Feed Reader software, although there will be some variations in the way an Atom feed entry is displayed when it contains multiple ‘link’ elements and a ‘content’ element. A prototype of a special (browser

AJAX) interface for reading, searching, and authoring servcasts has been developed (see http://sciflo.jpl.nasa.gov/wsgi/feed_reader, usr/pass = guest/guest). The prototype uses the ExtJS AJAX framework and some back-end python code. An example of this reader interface, populated with servcasts from ESIP Federation members and a few datacasts from JPL, is depicted in Figure 2.



SciFlo Data Query/Access Services
RGIS Base Imagery Service
ADaM Services: Data mining and imag
QuikSCAT Hurricane Ultra High-Res W
GHRSST AMSR-E SST
HD - NASA's Jet Propulsion Laborator
NSIDC Snow and Ice Cover

Title	Date
GeoRegionQuery Brian Wilson Space/time query and granule URL lookup services for multiple EOS satellite datasets at Levels L2/L3: AIRS, MODIS, MISR, GPS, and AERONET (ground network).	3/13 1:32 am
FindDataById Brian Wilson Translate a list of SciFlo data granule IDs to a list of on-line URLs.	3/13 1:32 am

GeoRegionQuery
by Brian Wilson
Mar 13, 2008, 1:32 am

Example Call

```
GeoRegionQuery(dataSetId='AIRS', level='L2', version=None, startTime='2006/01
/01T00:00:00', endTime='2006/02/01T00:00:00', latMin=-90., latMax=90., lonMin=-180.,
lonMax=180., responseGroups='Large' )
```

Service interface description · Server endpoint · Service documentation

Figure 2. Prototype servcast browser/reader with a catalog of servcasts on the left, a list of servcast entries from the selected SciFlo bundle on the top right, and details of the GeoRegionQuery service on the bottom right (including clickable links to SOAP/WSDL interface, endpoint, and docs.).

References

Atom syndication format, published as IETF RFC 4287.

Bingham, A., T. Stough, S. McCleese, R. G. Deen, K. Hussey, N. Toole (2008). "Earth Science Datacasting: Informed Pull and Information Integration," Accept for publication in *IEEE Transactions on Geoscience and Remote Sensing*, April, 2008.

Movva, S., et al., "Customizable Search Engine with Semantic and Resource Aggregation Capability," Proc. The Semantic Web meets the Deep Web Workshop, IEEE Joint Conference on E-Commerce Technology and Enterprise Computing, E-Commerce and E-Services 2008.

Ramachandran, R., et al., "Noesis: An Ontology-based Semantic Search Tool and Resource Aggregator," Proc. Geoinformatics 2006, 2006.

Smith, G. (2008). "Tagging: People-Powered Metadata for the Social Web", Berkeley, CA: New Riders. ISBN 0321529170.

Wilson, B. B. Tang, G. Manipon, D. Mazzoni, E. Fetzer, A. Eldering, A. Braverman, E. Dobinson, and T. Yunck, "GENESIS SciFlo: Scientific Knowledge Creation on the Grid Using a Semantically-Enabled Dataflow Execution Environment", peer-reviewed Proceedings of ***SSDBM 2005***.