Discovery White Paper Short

From Federation of Earth Science Information Partners

- Original Longer Version

Contents

- 1 Title: A Lightweight Approach to Earth Science Data Discovery
  - 1.1 Authors: The Earth Science Information Partners Discovery Cluster (ESIP-DC)
- 2 The Challenge and Vision
- 3 Requirements
- 4 The ESIP Federation and Discovery Cluster
- 5 CI Architecture Design, Development, and Integration
  - 5.1 Federated Search Framework
  - 5.2 Data and Service Casting Frameworks
  - 5.3 Beyond Discovery - Integrating Frameworks Together
  - 5.4 Technical Governance
- 6 Looking to the future
- 7 Discovery and Earth Science Collaboratory Cluster Participants

Title: A Lightweight Approach to Earth Science Data Discovery

Authors: The Earth Science Information Partners Discovery Cluster (ESIP-DC)

Points of Contact: Ruth Duerr, rduerr@nsidc.org, Chris Lynnes, christopher.s.lynnes@nasa.gov, Mark Parsons, parsonsm@nsidc.org, Chris A. Mattmann, chris.a.mattmann@nasa.gov, Hook Hua, hook.hua @ jpl.nasa.gov, Robert R. Downs, rdowns@ciesin.columbia.edu,

The Challenge and Vision

To truly understand the Earth system requires that we integrate increasingly diverse and complex data from all disciplines. But first one must find the data and see if they are useful. These seemingly simple first steps to data integration and workability remain significant challenges especially when seeking data across myriad disciplines, each with their own cultures of data collection, structure, and description. Traditional cataloging methods to enable discovery are reliant on cumbersome centralized registries, do not scale across disciplines, and do not allow adequate description of data to diverse audiences. The Federation of Earth Science Information Partners (ESIP) have, instead, developed a basic, grassroots, federated approach to discovery that does not rely on centralized registries. Further, we recognize that

data are often useless without associated services, so we also enable discovery of data services and associate those services with relevant data.

Requirements

The overarching requirement is for diverse audiences to be able to find data they need, discover data they did not initially realize they needed, and to assess the usefulness of the data. Data could come from any domain and consist of anything from a real time image streaming from a satellite to the in-depth knowledge of an Indigenous hunter.

ESIP partners come from many domains and serve very different users. A sub-group, the ESIP Discovery cluster assembled collective experience of use case development, usability testing, and interaction with different users and providers to create a generalized scenario that illustrates broad data and service discovery.

- An investigator starting a new project uses her favorite data portal to obtain a list of data and services meeting her spatial, temporal, and free text criteria. The portal conducts a search across a wide variety of services and data centers providing data that is of interest to the investigator. The results contain any advertised data set meeting her criteria no matter whether the data is held by an individual investigator or one of many data services around the world.
- The investigator examines several data sets and subscribes to be notified whenever new services or updates for those data sets become available.
- The investigator also subscribes to be notified when a new data set meeting her query criteria is available.
- The investigator finds results from two locations that are of interest. She retrieves the data from the first location using the link to the data set provided.
- While perusing service descriptions available for one of the data sets, the investigator sees that a useful specialized service is available. For example, she might discover a granule (e.g., file) level OpenSearch service produces KML output compatible with their favorite GIS analysis tool, so she uses that tool to browse the latest data, decide that the data is adequate, and download it all from within their analysis environment using a well-known protocol such as OPeNDAP (Open-Source Network for a Data Access Protocol) or Web Coverage Service.
- One morning, while perusing their feed reader, the investigator becomes aware that a new data set meeting her criteria has been published. They examine that data set and add it to their data set subscriptions.
- The investigator examines the services available for the new data set and is disappointed to find that no granule OpenSearch or equivalent service is available but that one is planned. They subscribe to receive service updates for the data set.
- Eventually the service she was interested in is released and they are automatically notified of that event.
- As a part of her work, the investigator generates a new data set that she decides should be published. She chooses to use an open-source web-based tool to announce availability of the data.
- Special purpose aggregation crawlers (for example one aggregating all geologic data) discover the new data cast while searching the web and add the data cast to their aggregations. Other investigators who subscribed to those aggregators are notified of the new data and may access the

data or contact the investigator.

The ESIP Federation and Discovery Cluster

The ESIP Federation has been addressing interworkable data since its inception in 1998. The federation is composed of a wide variety of Earth system science data, information and service providers: academic institutions, commercial interests, government agencies at federal, state and local levels, and non-governmental organizations. Members also cover a wide range of missions, from educational to research to applications, as well as a wide range of disciplines: solid-earth, oceanography, atmospheric sciences, land surface, ecology, and demographics.

This diversity has forced ESIP to confront many of the challenges to data integration. At the same time, it virtually mandates a loosely knit organization. While ESIP has a well-defined governance structure with respect to business activities, technical progress most often comes out of ESIP "clusters". Clusters are self-organizing groups of people within the federation who come together to tackle a particular issue, with integration across the Federation usually the main goal. Some clusters are domain-focussed, such as the Air Quality cluster, while others are formed to address particular aspects of information management, such as the Discovery, Preservation & Stewardship, and Semantic Web Clusters. More recently, a new Earth Science Collaboratory cluster has formed to help relate the diverse activities of ESIP and enhance scientific collaboration around data. Very much in the spirit of the EarthCube.

The Discovery Cluster began in 2009 and is the primary group working the broad data discovery and access issues described here. In keeping with the federated aspect of the ESIP Federation at large, a federated search solution was developed based on the OpenSearch (http://www.opensearch.org) conventions. In January of 2011, the Cluster began to include subscription based ("*casting") methods of discovery. The Discovery Cluster works to develop usable solutions to the problem of distributed and diverse providers, leveraging existing standards, conventions and technologies, with a predilection for simple solutions that have a high likelihood of voluntary adoption.

CI Architecture Design, Development, and Integration

Federated Search Framework

As described, the diversity of data and providers led us to favor a federated solution to the basic problem of Search. Federated search allows clients to search the contents of multiple, distributed catalogs simultaneously, merging the results for presentation to the user.Federated search for Earth science data has a long history. For example, the EOSDIS (Earth Observing System Data and Information System) Version 0 system implemented a federated search across eight remote sensing data centers as early as 1994, but the system did not scale well beyond the original core partners. A simpler solution was needed. In the business world, Amazon implemented a set of conventions to enable search across multiple vendors several years ago. The conventions, originally called A9 and later renamed OpenSearch (http://www.opensearch.org), have the virtue of being extremely simple to understand and implement. Simplicity is a critical aspect of any framework that needs widespread, voluntary adoption in a community with significant variation in technical capacity. Another advantage of OpenSearch is that it provides a way for data stewards to present information that would be otherwise inaccessible to search

engines and web crawlers. Most importantly, it allows those most familiar with the content in question to enable search in a way that is most appropriate to their holdings. Directly involving this subject knowledge view is likely to increase the relevance of search results.

Data and Service Casting Frameworks

Traditional federated search is a “pull” model for locating data where clients search sites known to have data of potential interest. Another model is a “publish and subscribe” model where providers simply advertise the existence of their data on their website using standard syndication protocols such as RSS and ATOM. These streams of data advertisements, called casts or feeds, are instantly discoverable, both by people who have subscribed and by standard web crawlers. Moreover, if the feeds are formatted using the OpenSearch protocols, then aggregation systems can find them and present them logically in federated search results. This approach allows investigators to easily expose their data to multiple client applications without significant technical support. They only need to generate a simple file and post it on their web site, and their data are advertised to multiple systems, even those unknown to the investigator. Similar mechanisms can be used to advertise changes or additions to a data set --additions to a time series, for example. It is even possible to automatically obtain the data upon receipt of the advertisement.

While finding data is challenging, finding tools and services to make effective use of the data can be even more challenging. Again, central registries have tried to address this, but the problems of complexity, burden to the provider, and poor interdisciplinary scalability are even greater with services than with data. Service casting, similar to data casting, provides a framework for allowing service producers to publish/advertise their services using a community-derived definition in standard RSS and ATOM protocols. The public availability of these feeds allow them to be discovered and/or subscribed to without using any specialized applications or tools. As with data, the service providers remain in control of their published service descriptions and can easily keep them up-to-date and complete.

A prototype project that provides information and tools for service casting is available at http://ws3dev.itsc.uah.edu/infocasting/.

Beyond Discovery - Integrating Frameworks Together

We have outlined several lightweight technologies:

- OpenSearch provides a simple mechanism for data stewards to describe and present data to the web in a way that is most appropriate for their data.
- Data Casting allows highly-distributed advertisement and subscription of data through simple popular protocols.
- Service Casting advertises services the same way as Data Casting.
- Aggregation services allow communities to discover and compile relevant data sources and then create tailored search and discovery interfaces.

Each of these technologies builds on simple well-established standards and protocols, and each by itself is useful. They can, however, be much more powerful when linked in a common framework. This, then, allows us to address the general scenario above. We find this approach of linking lightweight, targeted, simple components much more appropriate and scalable for distributed heterogeneous data than more

centralized systems relying on high-levels of technical agreement. Furthermore, our open, web-based approach not only allows search but also enables serendipitous discovery of previously unknown resources.

One critical issue that remains in integrating these technologies is to link data and services, so that users can readily discover tools to work with the data they find and vice versa. Addressing this issue in this lightweight framework is the current focus of the discovery cluster.

Technical Governance

The organization of the Discovery Cluster is volunteer based and consensus driven. Further, in keeping with our lightweight approach, we rely on open source software. Correspondingly, the expanding amount of open source software and licenses, software management models and providers, IP issues, and community interaction models highlights a need for an “open source understanding framework”. Members of the Cluster have been developing this framework and have been presenting it to NASA, ESIP, and NSFs NEON project. We have suggested a community-governed model for NASA to follow for consuming and producing open source software, have identified strategic dimensions and tradeoffs and have developed an initial set of recommendations to move forward.

In an interoperability regime involving multiple government agencies at all levels, academia, the commercial world and even citizen scientists, it is difficult for any single organization or group to set forth interoperability standards for all to follow. Instead, the EarthCube governance should allow for emergent governance structures that can leverage the efforts of both volunteer and funded participants. The cluster formation within the ESIP federation provides an example of how this can work.

With multiple data centers from a variety of different organizations now involved with the Discovery Cluster activities, resolving issues as a virtual organization becomes increasingly difficult. The Cluster routinely collaborates to resolve issues related to interoperability, distributed services, and adoption of different open standards. While ESIP is not a standards body, a structured process is still needed to coordinate the various viewpoints, decisions, and interoperability issues. Therefore developing an agreed upon process, and then following it, becomes critical to the ongoing collaboration of the various organizations that compose the Cluster.

Given the agile nature of the grass roots-like developments of Discovery Cluster’s lightweight approach to Earth science data discovery, we needed a governance process that was also lightweight and agile. In 2010, the Cluster adopted a governance process borrowing useful ideas from the Open Provenance Model governance (http://twiki.ipaw.inføpub/OPM/WebHome/governance.pdf) process, which had similar lightweight and agile needs. The Cluster’s governance process encompasses the following steps:

1. Submission of new proposals
2. Forum to review proposals
3. Author revision based on feedback
4. Voting on change proposals
5. Ratification or rejection by editors

To maintain an open community process, all steps are posted to the mailing list and/or wiki.

Looking to the future

The ESIP Discovery Cluster has begun to demonstrate how a lightweight standard or convention can enable significant discoverability of data and services and how similar, interlocking conventions can provide cross-cutting interoperability between services and data. We see this as a first step in our drive to make systems "interworkable". Data and services (or tools) can be combined in sequences to form scientific workflows. The analysis results from executing these workflows may also be thought of in a fashion similar to data. And the results themselves may be aggregated into an experiment, in much the same way that different model runs are aggregated into an ensemble. Many of the key discovery attributes of workflows, results and experiments can be inherited from the data and service building blocks from which they are made. As a result, it is not too ambitious to hope that the entire "information stack", from data and services, up through workflows, results and experiments, can be interoperable (or interworkable) both horizontally (data with data, result with result) and vertically (data with tool with workflow with result with experiment).

This interoperability framework has the key advantage of presenting everything in the proper context--a given result could be traced back down through the analysis workflow to the services and data that led to the result. We seek then to further enhance this rich context with basic social networking technology, allowing researchers to annotate any level of the information stack with expert or contextual knowledge. Such an "Earth Science Collaboratory (ESC)" (Fig. 1) has been proposed within ESIP, with an Earth Science Collaboratory Cluster formed to push the idea forward.

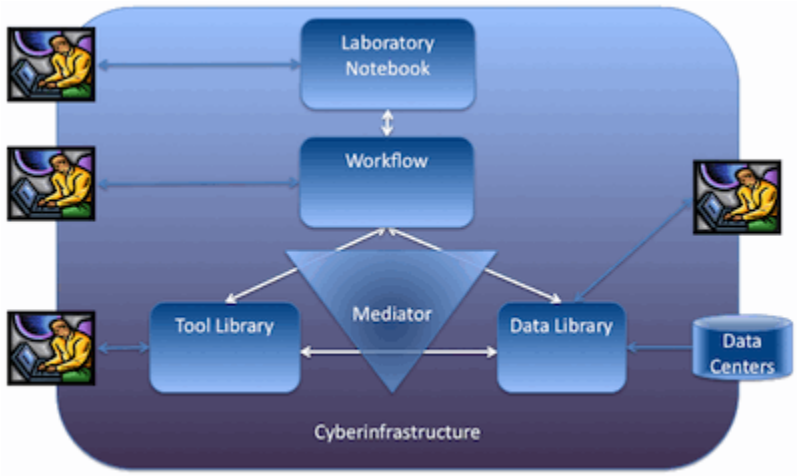

Fig.1 Architectural concept for an Earth Science Collaboratory (ESC). Key repositories for data, tools, workflows and analysis results enable rich, contextual sharing within the community.

The Collaboratory would allow researchers to share not just data, but tools, services, analysis workflows (i.e., techniques), and results as easily as links are shared today in tools such as Facebook, thus capturing provenance and preserving the full context of a given result as well as the contextual knowledge added by the researcher. This could have further potential benefits for many other types of user. For instance, science assessment committees would be able to share with each other both the (usually highly processed) end results and articles but also the input data and tools, greatly increasing transparency of the assessment. Novice graduate students would be able to "follow" more experienced researchers in the

field, thus learning how to handle the data properly and avoiding common pitfalls. Educators would be able to put together science stories that trace back to the original data, allowing them to give students exposure to what "real" data look like, and how they are eventually processed to yield a compelling story. Users of Decision Support Systems (DSS) would be able to collaborate in real time with the scientist whose research is incorporated into the DSS, providing a valuable bridge over the chasm that often separates research and operations.

Creating the ESC still faces a number of social and technical challenges, but ESIP has laid down an initial foundation of lightweight governance and technologies that provide significant capability while allowing great flexibility. The NSF EarthCube vision aligns closely with our efforts and could, therefore, provide the critical impetus toward realization of a fully interworkable Earth Science Collaboratory.

Discovery and Earth Science Collaboratory Cluster Participants

- Clynnes 10:35, 15 September 2011 (MDT)
- Kskuo 13:04, 15 September 2011 (MDT)
- Hook 15:37, 15 September 2011 (MDT)
- Ctilmes 10:14, 19 September 2011 (MDT)
- Keiser 08:57, 20 September 2011 (MDT)
- Brianwee 10:43, 20 September 2011 (MDT)
- Parsonsm 13:47, 6 October 2011 (MDT)
- Rduerr 13:36, 9 October 2011 (MDT)
- Rdowns 14:13, 11 October 2011 (MDT)