ESIP Federated Search for Water Cycle Data: Cluster Mashup

Christopher Lynnes, Goddard Earth Sciences Data and Information Sciences Center
Hook Hua, ACCESS-NEWS

Finding and obtaining data across a diverse set of providers remains a tedious, time-consuming process. A plethora of dataset directories exist, yet searching for specific data files often requires going to individual provider search tools. The ESIP Federated Search Cluster was formed to provide a low-cost, grassroots approach to this problem. Water Cycle data in particular is susceptible to this issue, with key data resources dispersed at a wide variety of sites. We will present the progress of a project to implement end-to-end federated search across multiple Water Cycle data providers, beginning first with the GES DISC and the ACCESS-NEWS projects, but eventually pulling in such sources as the Global Hydrology Research DISC and the ACCESS-NEWS projects, but eventually Water Cycle data providers, beginning first with the GES DISC and the ACCESS-NEWS projects, but eventually.

Abstract
Finding and obtaining data across a diverse set of providers remains a tedious, time-consuming process. A plethora of dataset directories exist, yet searching for specific data files often requires going to individual provider search tools. The ESIP Federated Search Cluster was formed to provide a low-cost, grassroots approach to this problem. Water Cycle data in particular is susceptible to this issue, with key data resources dispersed at a wide variety of sites. We will present the progress of a project to implement end-to-end federated search across multiple Water Cycle data providers, beginning first with the GES DISC and the ACCESS-NEWS projects, but eventually pulling in such sources as the Global Hydrology Research DISC and the ACCESS-NEWS projects, but eventually.

Water Cycle data are stored all over the map

Step 1: Dataset Discovery
Search for datasets and select the ones of interest for granule-level search

Step 2: Granule Search
For selected datasets, execute space-time query for desired granules

The Two-Step process avoids the effort of searching and weeding through granule results for irrelevant datasets.

Key Challenge: how to use the results from Step 1 in Step 2
- What host do we query for granules?
- What is the syntax for the granule-level query?
- What kind of dataset identifier do we pass to the query?

Answer: Recursive OpenSearch
- Dataset-level results link to the OpenSearch Description Documents for the granule-level query
- Dataset identifier is already filled in the template, e.g.:

```
<os:Url type="application/atom+xml"
  template="http://mirador.gsfc.nasa.gov/cgi-bin/mirador/granlist.pl?
  dataSet=AIRS2RET.005"&dataSet=MergedWaterCycle&
  pointLocation={geo:box}&endPoint={time:end}&
  &format=atom"&maxgranules={count}&
  &startTime={time:start}&
  &endTime={time:end}&
  &page=1&maxgranules={count}&
  &format=atom>
```

Just replace placeholders with search criteria and fetch the URL.

What do I need for an ESIP Federated Search server?
- Supports OpenSearch keyword query
- Supports OpenSearch with Time and Space Extensions
- Return Atom documents following ESIP Federated Search recommendations for <link> elements
- OpenSearch Description Documents for granule search engine
- Describe the syntax of the Granule Search Engine, for each dataset

OpenSearch Description Document
- Describes dataset-level query
- Returns Atom documents with links to OpenSearch Description Documents for Granule Search Engine

Servers in Development
- ACCESS-NEWS
- ECHO
- GES DISC
- MODAPS
- NSIDC

ESIP Federated Search Clients
A client can be as simple as an XSLT, attached to the OpenSearch Description Document:

```
<os:Url type="application/atom+xml"
  template="http://mirador.gsfc.nasa.gov/cgi-bin/mirador/granlist.pl?
  dataSet=AIRS2RET.005"&dataSet=MergedWaterCycle&
  pointLocation={geo:box}&endPoint={time:end}&
  &format=atom"&maxgranules={count}&
  &startTime={time:start}&
  &endTime={time:end}&
  &page=1&maxgranules={count}&
  &format=atom>
```

Also available: a reference implementation / test script
~80 lines of simple Perl

Clients in Development
- Mirador (GES DISC): a granule-level search client
- Talkoot (UAH): an Earth science collaboration framework

An initial end-to-end attempt: Mirador + ACCESS-NEWS

Federated Search: If a client could search all the provider sites, the way Amazon searches its affiliated providers for that obscure Vikram Seth book...

Wikipedia: “A Federation is multiple computing and/or network providers agreeing upon standards of operation in a collective fashion.”

The ESIP Federated Search Cluster was formed to agree upon standards for distributed search of participating providers:
- Should support search down to the granule / file level
- Should support space-time query
- http://wiki.esipfed.org/index.php/Federated_Search

ESIP Federated Search is based on OpenSearch conventions