How To Share Data With Your Collaborators

Shreyas Cholia
Data and Analytics Services, NERSC

Joint Facilities User Forum on Data-Intensive Computing, Oakland, June 18, 2014
Overview

• Introduction
• Data Sharing over HTTP
• APIs
• High Performance Data Access
• Data Management Systems
• We’ll try to do a few live demos along the way
Introduction
Data Islands

• Data is still heavily siloed and inaccessible.
• Data sits on a machine somewhere, and you give people local accounts to access it.
• Good luck combining multiple datasets
• Does not scale!

• This is 2014 – we can do better!
• **OSTP requirements around data management**
  
  – All proposals submitted to the Office of Science for research funding are required to include a Data Management Plan (DMP).
  
  – **DMPs must provide a plan for making all research data displayed in publications resulting from the proposed research digitally accessible at the time of publication.**
  
  – Researchers that plan to work at an Office of Science User Facility as part of the proposed research should consult the published data policy of that facility.
J.M. Wicherts, M. Bakker, and D. Molenaar: Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results


*Content for this slide courtesy Greg Wilson, Software Carpentry*
Data Sharing

- Common access to data enables collaboration
- Opens up the possibility of cross comparisons, combining multiple datasets from different sources
- Enables new science use cases
Data Sharing over HTTP
Sharing Data Over The Web

• Provide projects with a landing spot for data within the center boundary
  – High performance access to filesystems, backend resources etc.

• Serve up landing spot via web server
  – Minimally provides a very simple publication sharing system
Data Sharing Over The Web

- Store data on persistent global filesystem
- Designate an area for export
- Mount export area RO on a web node using a shared FS
- Run a webserver on web node to share with world
- Add AuthNZ to web server configuration
- Demo time!
Authorization

• **Authorization:** Access privileges for a given identity
  – What is the user allowed to do?

• **Data portals see some combination of the following sharing modes:**
  – Public – data are publicly viewable to all users
  – Protected – data are only visible to members within a designated group or collaboration
  – Private – data are private to the individual user
APIs
• Make your data available as a Web API to enable more fine-grained queries
• Use HTTP Method + URLs + parameters to query and access data
• Users just download what they need, server does heavy lifting in terms of search and sub-selection
  – eg. Google maps - you don't download the map just the tiles you are interested in.
• Enable rich interactions and applications around your data
• eg. OpenDAP connects NETCDF/HDF5 datasets with web clients
  – Demo: [http://portal.nersc.gov/pydap](http://portal.nersc.gov/pydap)
Data Browsers

• HTML5/JavaScript + Server Side APIs to build data browsers and viewers
• Make callouts to REST API for all server-side information
• eg. https://openmsi.nersc.gov
Getting started – Hello World API

```python
from flask import Flask
from flask import jsonify
app = Flask(__name__)

# http://localhost:5000/object_id - returns object
@app.route('/<object_id>')
def get_data(object_id):
    obj = get_object_from_db(object_id)
    return jsonify(obj)

# http://localhost:5000/ - returns all objects
@app.route('/')
def get_all_data():
    obj_list = get_data_from_db()
    return jsonify(obj_list)

if __name__ == "__main__":
    app.run()
```
High Performance Data Sharing
High Performance Data Sharing

- Globus
- webdav + xrootd
- Aspera (Commercial Tool)
Globus Sharing

• Enable high performance GridFTP access to your data
• Share with designated groups or individuals
• https://www.globus.org/data-sharing
Data Management Portals
Data Management Systems (DMS)

- Data Management System (like a CMS but with a focus on Data)
- CKAN – open source tool for data publishing & management used by data.gov.uk and publicdata.eu
  - Publish and Manage Data
  - Search and Discovery
  - Metadata
  - Geospatial and Visualization Features
  - Community
  - Link to Filesets or store within portal
  - Data Provenance
  - Extensibility and APIs
  - Federation
CKAN demo

• http://ckan.org
Federated Data

• In some cases data may be distributed across multiple locations
• There are federated portals like ESGF which enable searching across data from multiple locations
• Metadata is federated but data is ultimately downloaded from source portal
ESGF – Federated Search

Current Selections
- (x) query: cmor

Search Categories
- Project
- Institute
- Model
- SubModel
- Instrument
- Experiment Family
- Experiment
- SubExperiment
- Time Frequency
- Product
- Realm
- Variable
- Variable Long Name

Search All Sites  Show All Replicas  Show All Versions

To download data: add datasets to your Data Cart, then click on Expand or wget.

< 1 2 3 ... 6518 6519 >  displaying 1 to 10 of 65186 search results

Display 10 datasets per page

Add All Displayed to Data Cart  Remove All Displayed from Data Cart

Results

project=CMIP5, model=EC-EARTH consortium, experiment=10- or 30-year run initialized in year 1960, time_frequency=day, modeling_realm=sealce, ensemble=r1i3p1, version=20140318

Data Node: esg-dn1.nsc.liu.se

Version: 20140318
Description: EC-EARTH model output prepared for CMIP5 10- or 30-year run initialized in year 1960
Further options: Add To Cart  Visualize and Analyze  Model Documentation

project=CMIP5, model=EC-EARTH consortium, experiment=10- or 30-year run initialized in year 1960, time_frequency=day, modeling_realm=sealce, ensemble=r1i3p1, version=20140318

Data Node: esg-dn1.nsc.liu.se

Version: 20140318
Description: EC-EARTH model output prepared for CMIP5 10- or 30-year run initialized in year 1960
Further options: Add To Cart  Visualize and Analyze  Model Documentation
The End

• Contact:
  – Shreyas Cholia: scholia@lbl.gov

• Questions? Comments?