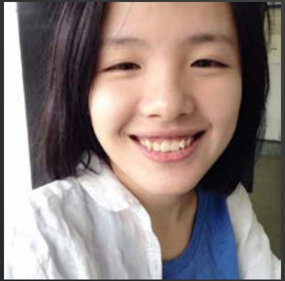
Building on the NOAA Big Data Project for Academic Research: An OCC Perspective

Zachary Flamig

Maria Patterson

Yajing (Phillis)Tang



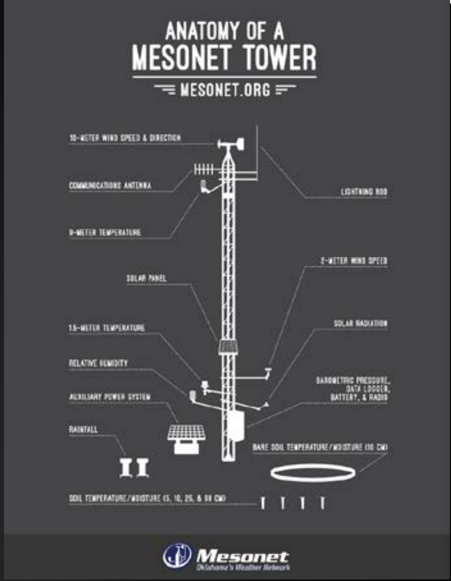
Walt Wells

Robert Grossman

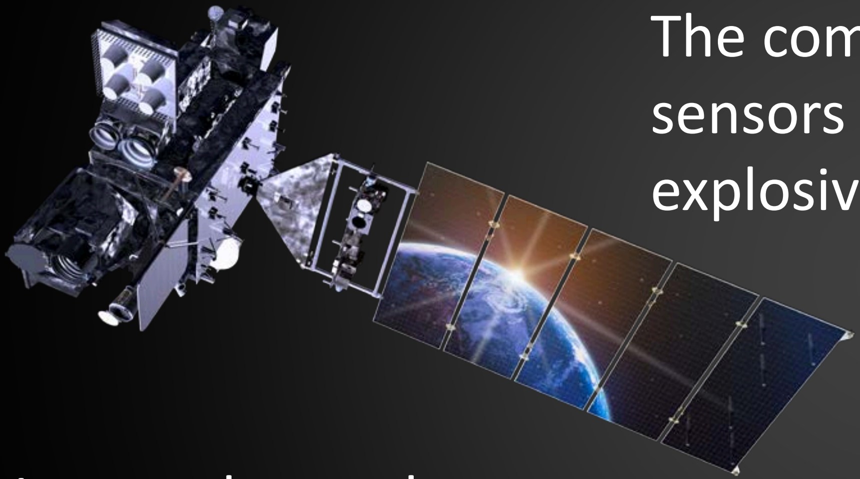
THE UNIVERSITY OF CHICAGO | Center for Data Intensive Science

OPEN COMMONS CONSORTIUM

We have a problem…



The commoditization of sensors is creating an explosive growth of data.



It can take weeks to download large datasets.

There is not enough funding for every researcher to house all the data they need.

Analyzing the data is more expensive than producing it.

Data Commons

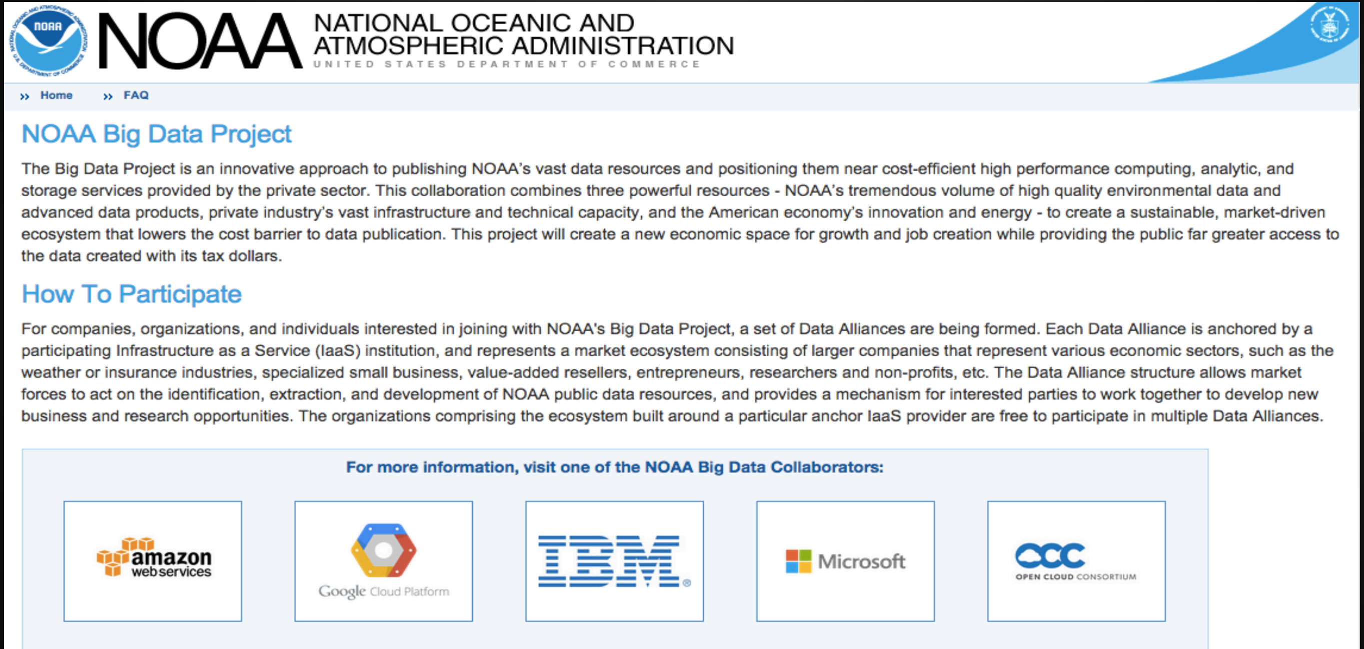
Data commons co-locate data, storage and computing infrastructure, and commonly used tools for analyzing and sharing data to create a resource for the research community.

NIH NATIONAL CANCER INSTITUTE
Genomic Data Commons

datacommons@psu
data: energy.environment.society.technology

A Case for Data Commons: Toward Data Science as a Service
Grossman, Robert L. and Heath, Allison and Murphy, Mark and Patterson, Maria and Wells, Walt,
Computing in Science & Engineering, **18**, 10-20 (2016), DOI: 10.1109/MCSE.2016.92

NOAA Big Data Project



NOAA
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
UNITED STATES DEPARTMENT OF COMMERCE
» Home
» FAQ

NOAA Big Data Project

The Big Data Project is an innovative approach to publishing NOAA's vast data resources and positioning them near cost-efficient high performance computing, analytic, and storage services provided by the private sector. This collaboration combines three powerful resources - NOAA's tremendous volume of high quality environmental data and advanced data products, private industry's vast infrastructure and technical capacity, and the American economy's innovation and energy - to create a sustainable, market-driven ecosystem that lowers the cost barrier to data publication. This project will create a new economic space for growth and job creation while providing the public far greater access to the data created with its tax dollars.

How To Participate

For companies, organizations, and individuals interested in joining with NOAA's Big Data Project, a set of Data Alliances are being formed. Each Data Alliance is anchored by a participating Infrastructure as a Service (IaaS) institution, and represents a market ecosystem consisting of larger companies that represent various economic sectors, such as the weather or insurance industries, specialized small business, value-added resellers, entrepreneurs, researchers and non-profits, etc. The Data Alliance structure allows market forces to act on the identification, extraction, and development of NOAA public data resources, and provides a mechanism for interested parties to work together to develop new business and research opportunities. The organizations comprising the ecosystem built around a particular anchor IaaS provider are free to participate in multiple Data Alliances.

For more information, visit one of the NOAA Big Data Collaborators:

Public-private data collaborative announced April 21, 2015 by Secretary of Commerce Pritzker.

OCC Point of View

- *The research community* and NOAA Data Alliance working group will help determine

  1) which datasets benefit the community most by being placed in the cloud?
  2) which corresponding tools are the most useful for working with these data?
  3) how can we implement ID and metadata services for finding/linking data of interest?

- We work with NOAA to place selected datasets in the cloud and make them available to the community at no cost.

- We provide and enable value added services of interest over these data to the NOAA research community.

Working Group Leadership

UCAR COMMUNITY PROGRAMS | unidata
Data Services and Tools for Geoscience

The HDF Group

NCSA

THE UNIVERSITY OF CHICAGO | Center for Data Intensive Science

NOAA Data Available

- 2015 NEXRAD Level 2 Data – All Radars
- Real-time GOES-13 & -15 feed
  - 7TB rolling archive
- CFS Reanalysis 1979-2011
- Storm Data 1979-Present
- VIIRS
  - Day/Night Band so far
  - Bands 1,2,3,4 coming soon
  - 200TB rolling archive through August

NOAA Data Coming Soon

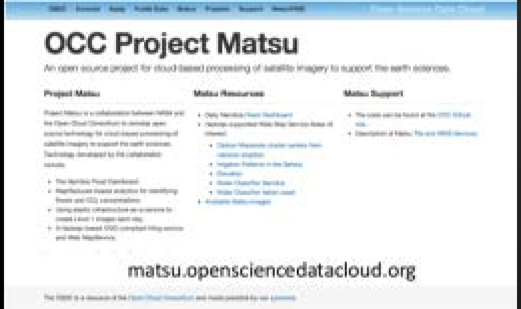
- GOES-16, Next week?
  - 100TB rolling archive
- National Water Model
  - Reanalysis ~ 40TB
- What data would you like to see here?
  - METARs? Sounding Archive? Text Products?
  - Fish genomics? Ocean data?
  - Model Archives? CFS forecast archive?

NOAA Data Coming Soon

- What data would you like to see here?
  - Global Ocean Ship-Based Hydrographic Investigation Program (GO-SHIP)
  - Gpsmet TPW
  - ARGO Float Profiles
  - Hurricane Research Data
  - UAS Survey Data
  - Global Surface Drifters
  - HRRR archive from GSD
  - National Energy Weather System
  - Wind Forecast Improvement Study

OSDC OPEN SCIENCE DATA CLOUD

- 850+ research projects supported since 2010.
- Over 20 million core hours used by allocation grantees in past year
- OSDC Griffin: 610 cores, 470TiB, Openstack w/ Ceph Object Storage



OCC Project Matsu

An open source project for cloud-based processing of satellite imagery to support the earth sciences.

Project Matsu

[illegible]

Matsu Resources

[illegible]

Matsu Support

[illegible]

matsu.opensciencedatacloud.org

BIONIMBUS PROTECTED DATA CLOUD

Secure cloud services for the scientific community

What is the Bionimbus PDC?

The Bionimbus Protected Data Cloud (PDC) is a collaboration between the Open Science Data Cloud (OSDC) and the IGSB (IGSB), the Center for Research Informatics (CRI), the Institute for Translational Medicine (ITM), and the University of Chicago Comprehensive Cancer Center (UCCCC). The PDC allows users authorized by NIH to compute over human genomic data from dbGaP in a secure compliant fashion. Currently, selected datasets from the The Cancer Genome Atlas (TCGA) are available in the PDC.

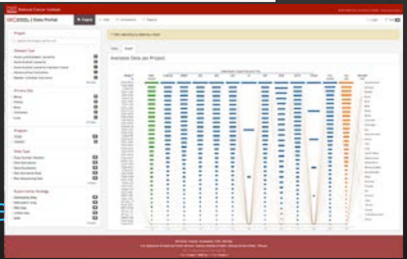
How can I get involved?

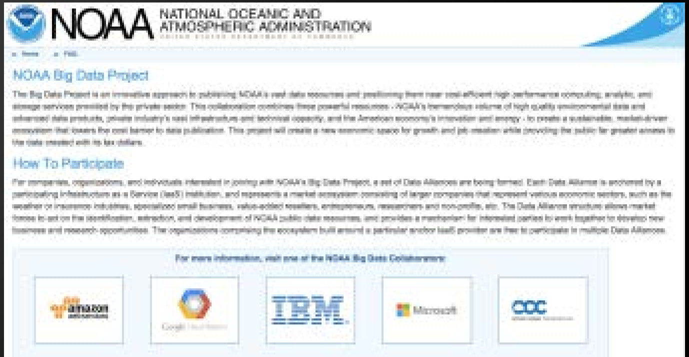
- Apply for an Bionimbus PDC account and use the Bionimbus PDC to manage, analyze and share your data.
- Partner with us and add your own racks to the Bionimbus PDC (we will manage them for you).
- Help us develop the open source Bionimbus PDC software stack.

You can contact us at info@opencloudconsortium.org.

How do I get started?

First, apply for an account. Once your account is approved, you can login to the console and get started. Support questions can be directed to support@opencloudconsortium.org.





NOAA
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
UNITED STATES DEPARTMENT OF COMMERCE

NOAA Big Data Project

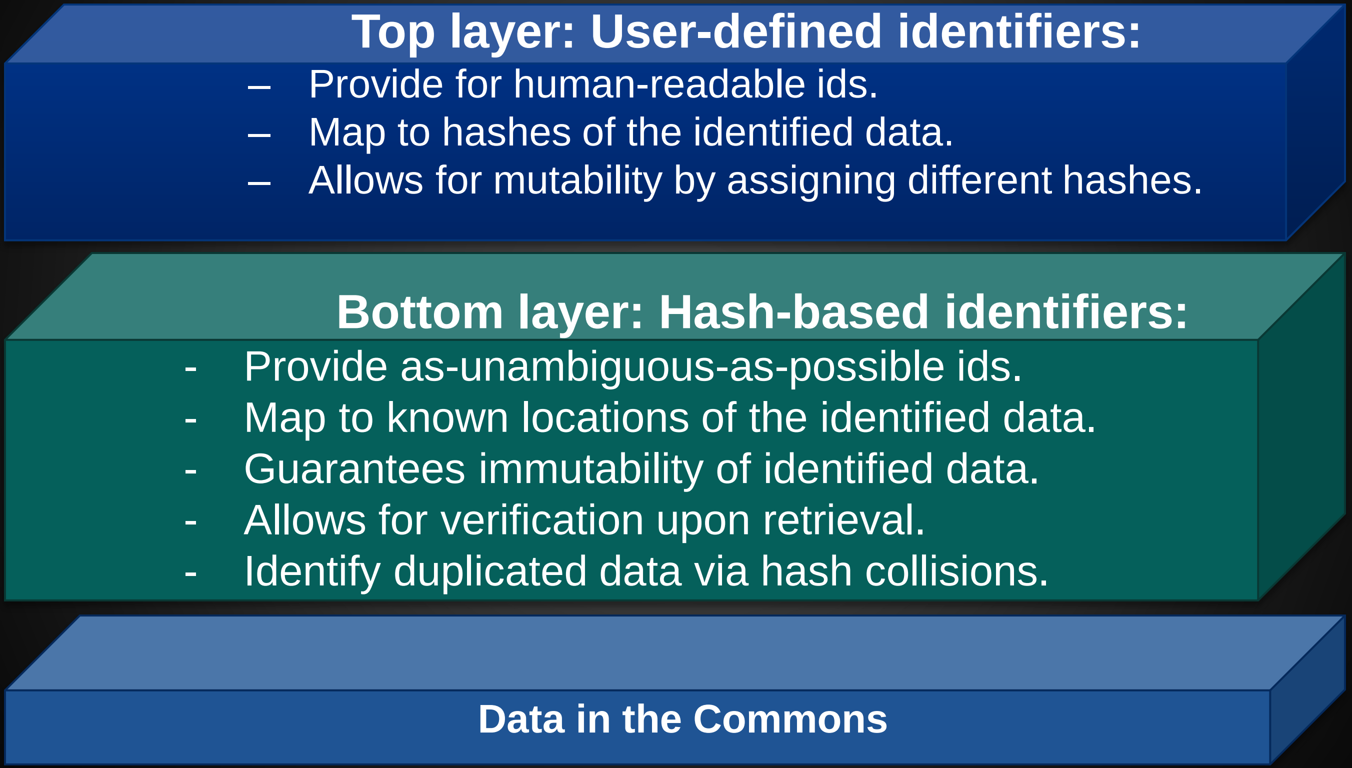
The Big Data Project is an innovative approach to publishing NOAA's vast data resources and positioning them near cost-efficient high performance computing, analytic, and storage services provided by the private sector. This collaboration combines three powerful resources - NOAA's tremendous volume of high quality environmental data and advanced data products, private industry's vast infrastructure and technical capacity, and the American economy's innovation and energy - to create a sustainable, market-driven ecosystem that lowers the cost barrier to data publication. This project will create a new economic space for growth and job creation while providing the public far greater access to the data created with its tax dollars.

How To Participate

For companies, organizations, and individuals interested in joining with NOAA's Big Data Project, a set of Data Alliances are being formed. Each Data Alliance is anchored by a participating Infrastructure as a Service (IaaS) institution, and represents a market ecosystem consisting of larger companies that represent various economic sectors, such as the weather or insurance industries, specialized small business, value-added resellers, entrepreneurs, researchers and non-profits, etc. The Data Alliance structure allows market forces to act on the identification, extraction, and development of NOAA public data resources, and provides a mechanism for interested parties to work together to develop new business and research opportunities. The organizations comprising the ecosystem built around a particular anchor IaaS provider are free to participate in multiple Data Alliances.

For more information, visit one of the NOAA Big Data Collaborators:

Data Commons: Enabled with ID services



Top layer: User-defined identifiers:

- Provide for human-readable ids.
- Map to hashes of the identified data.
- Allows for mutability by assigning different hashes.

Bottom layer: Hash-based identifiers:

- Provide as-unambiguous-as-possible ids.
- Map to known locations of the identified data.
- Guarantees immutability of identified data.
- Allows for verification upon retrieval.
- Identify duplicated data via hash collisions.

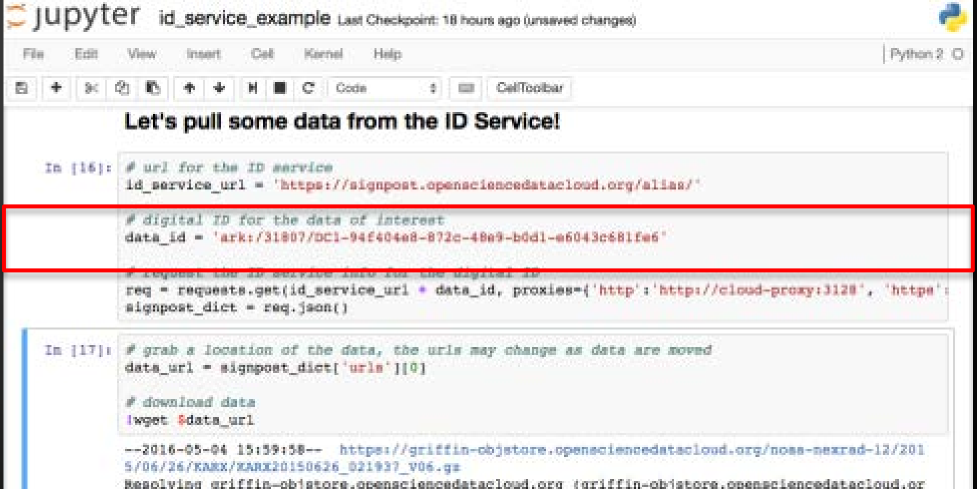
Data in the Commons

Finding data using ID service query tool

Researcher selects search parameters like weather station and start and end date.

Query tool returns relevant digital IDs.

Researcher can then use digital IDs to download data directly or to reference data in their analysis scripts.

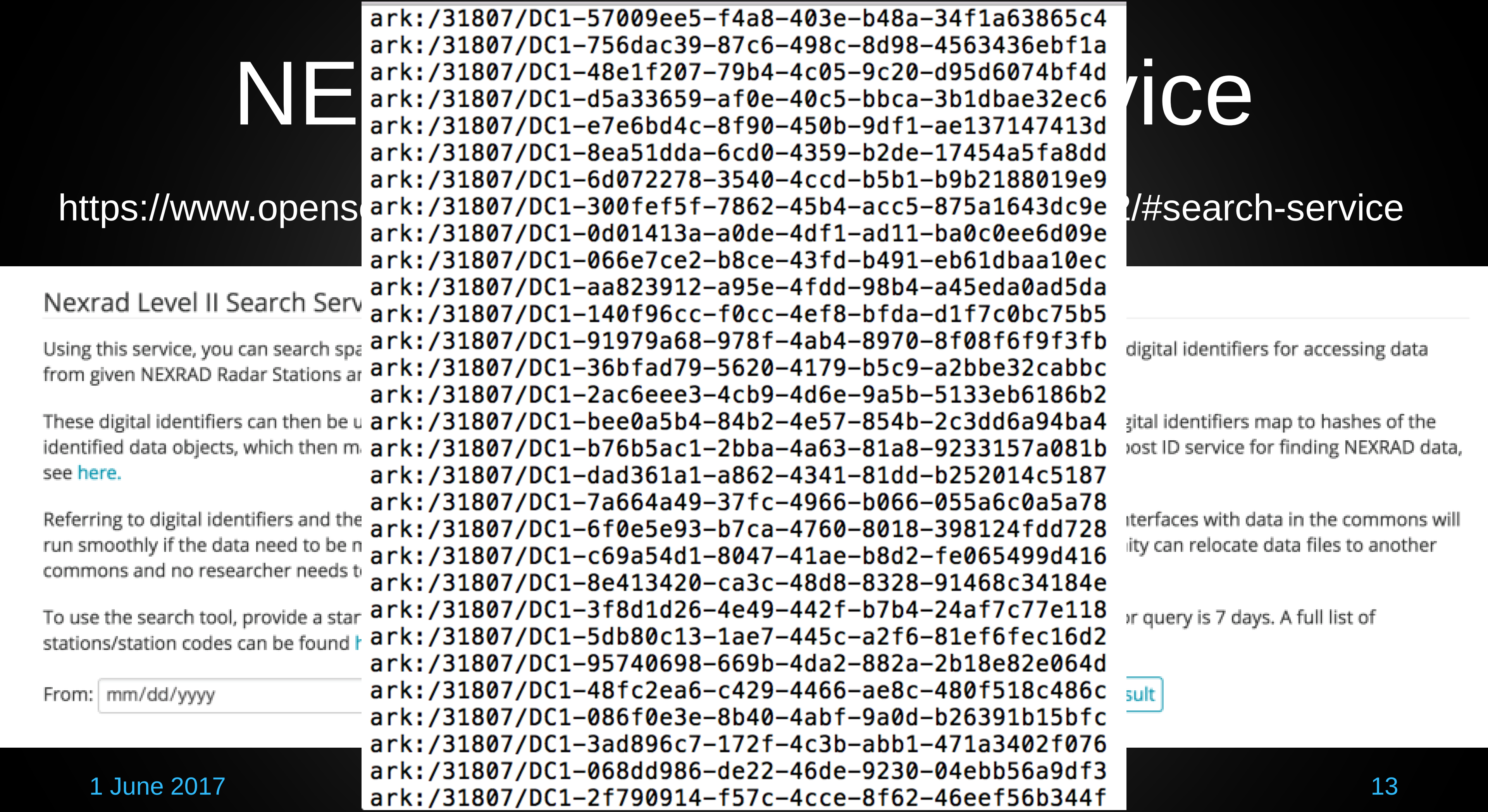

jupyter
id_service_example Last Checkpoint: 18 hours ago (unsaved changes)
File
Edit
View
Insert
Cell
Kernel
Help
Python 2
Code
CellToolbar

Let's pull some data from the ID Service!

```
In [16]: # url for the ID service
         id_service_url = 'https://signpost.opensciencedatacloud.org/alias/'

         # digital ID for the data of interest
         data_id = 'ark:/31807/DC1-94f404e8-872c-48e9-b0d1-e6043c681fe6'

         # request the ID service info for the digital ID
         req = requests.get(id_service_url + data_id, proxies={'http':'http://cloud-proxy:3128', 'https':
         signpost_dict = req.json()

In [17]: # grab a location of the data, the urls may change as data are moved
         data_url = signpost_dict['urls'][0]

         # download data
         !wget $data_url

         --2016-05-04 15:59:58--  https://griffin-objstore.opensciencedatacloud.org/noaa-nexrad-l2/201
         5/06/26/KARX/KARX20150626_021937_V06.gz
         Resolving griffin-objstore.opensciencedatacloud.org (griffin-objstore.opensciencedatacloud.or
```

NE vice

https://www.opens /#search-service



Nexrad Level II Search Serv

Using this service, you can search spa digital identifiers for accessing data
from given NEXRAD Radar Stations ar

These digital identifiers can then be u gital identifiers map to hashes of the
identified data objects, which then m ost ID service for finding NEXRAD data,
see here.

Referring to digital identifiers and the terfaces with data in the commons will
run smoothly if the data need to be n ity can relocate data files to another
commons and no researcher needs t

To use the search tool, provide a star r query is 7 days. A full list of
stations/station codes can be found

From: mm/dd/yyyy

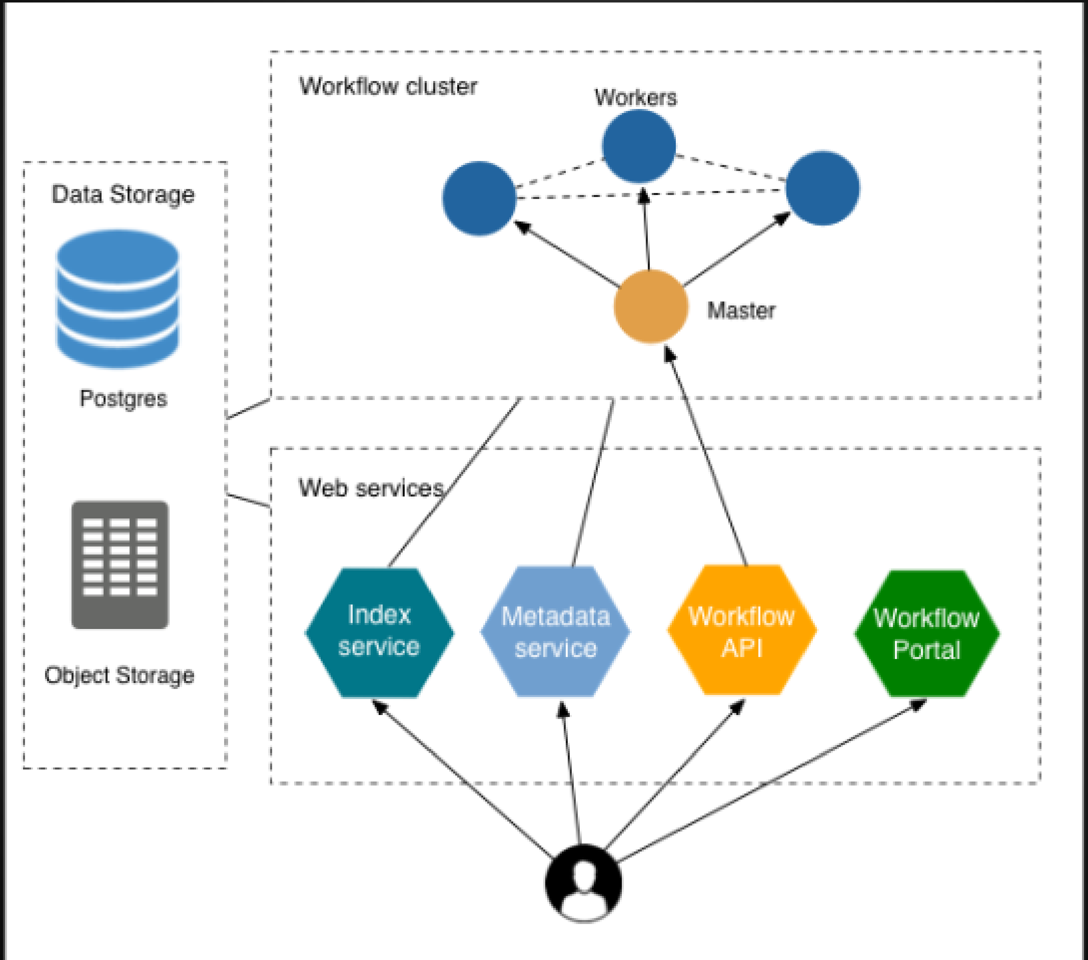
sult

```
ark:/31807/DC1-57009ee5-f4a8-403e-b48a-34f1a63865c4
ark:/31807/DC1-756dac39-87c6-498c-8d98-4563436ebf1a
ark:/31807/DC1-48e1f207-79b4-4c05-9c20-d95d6074bf4d
ark:/31807/DC1-d5a33659-af0e-40c5-bbca-3b1dbae32ec6
ark:/31807/DC1-e7e6bd4c-8f90-450b-9df1-ae137147413d
ark:/31807/DC1-8ea51dda-6cd0-4359-b2de-17454a5fa8dd
ark:/31807/DC1-6d072278-3540-4ccd-b5b1-b9b2188019e9
ark:/31807/DC1-300fef5f-7862-45b4-acc5-875a1643dc9e
ark:/31807/DC1-0d01413a-a0de-4df1-ad11-ba0c0ee6d09e
ark:/31807/DC1-066e7ce2-b8ce-43fd-b491-eb61dbaa10ec
ark:/31807/DC1-aa823912-a95e-4fdd-98b4-a45eda0ad5da
ark:/31807/DC1-140f96cc-f0cc-4ef8-bfda-d1f7c0bc75b5
ark:/31807/DC1-91979a68-978f-4ab4-8970-8f08f6f9f3fb
ark:/31807/DC1-36bfad79-5620-4179-b5c9-a2bbe32cabbc
ark:/31807/DC1-2ac6eee3-4cb9-4d6e-9a5b-5133eb6186b2
ark:/31807/DC1-bee0a5b4-84b2-4e57-854b-2c3dd6a94ba4
ark:/31807/DC1-b76b5ac1-2bba-4a63-81a8-9233157a081b
ark:/31807/DC1-dad361a1-a862-4341-81dd-b252014c5187
ark:/31807/DC1-7a664a49-37fc-4966-b066-055a6c0a5a78
ark:/31807/DC1-6f0e5e93-b7ca-4760-8018-398124fdd728
ark:/31807/DC1-c69a54d1-8047-41ae-b8d2-fe065499d416
ark:/31807/DC1-8e413420-ca3c-48d8-8328-91468c34184e
ark:/31807/DC1-3f8d1d26-4e49-442f-b7b4-24af7c77e118
ark:/31807/DC1-5db80c13-1ae7-445c-a2f6-81ef6fec16d2
ark:/31807/DC1-95740698-669b-4da2-882a-2b18e82e064d
ark:/31807/DC1-48fc2ea6-c429-4466-ae8c-480f518c486c
ark:/31807/DC1-086f0e3e-8b40-4abf-9a0d-b26391b15bfc
ark:/31807/DC1-3ad896c7-172f-4c3b-abb1-471a3402f076
ark:/31807/DC1-068dd986-de22-46de-9230-04ebb56a9df3
ark:/31807/DC1-2f790914-f57c-4cce-8f62-46eef56b344f
```

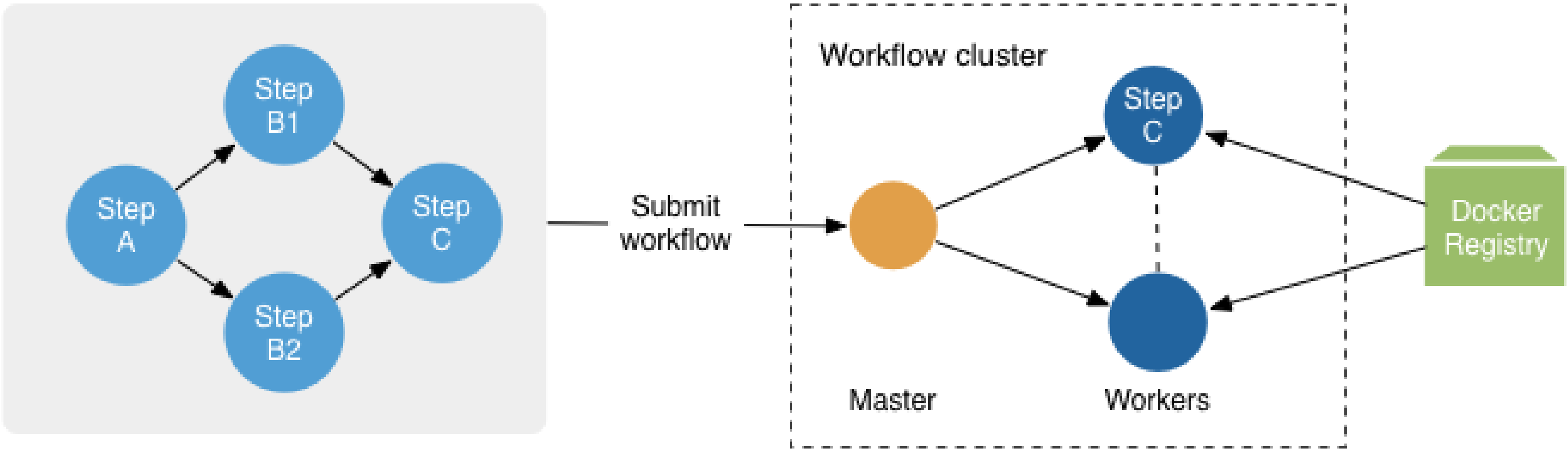
Denver Commons

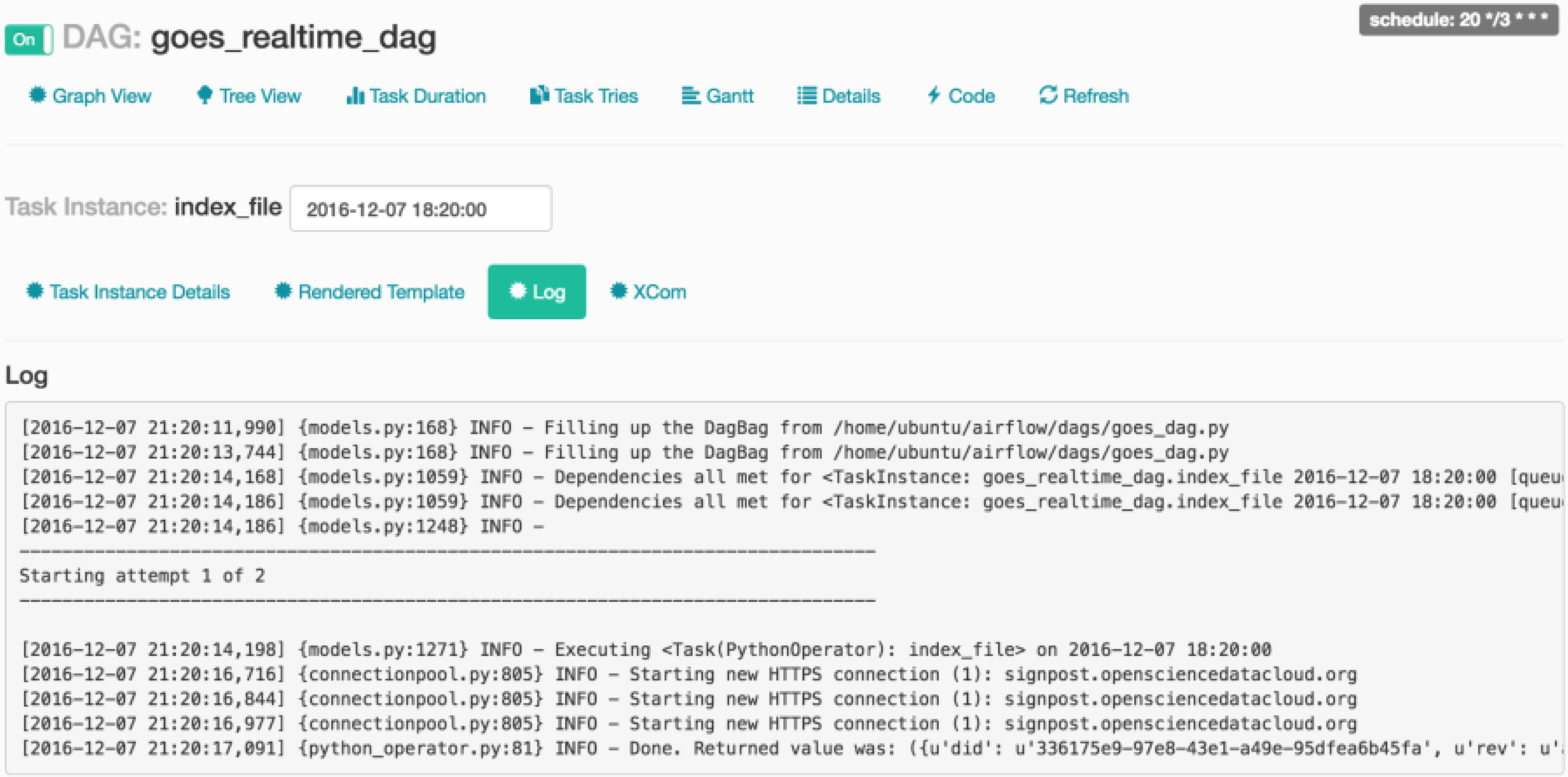
- First attempt at putting together storage, indexing, compute, and workflow management for disparate datasets
- Using Airflow developed by Airbnb, now under Apache incubation. Also utilizing Celery, Consul, and Docker

Denver Commons



Denver Commons





Task Instance: index_file 2016-12-07 18:20:00

Log

```
[2016-12-07 21:20:11,990] {models.py:168} INFO - Filling up the DagBag from /home/ubuntu/airflow/dags/goes_dag.py
[2016-12-07 21:20:13,744] {models.py:168} INFO - Filling up the DagBag from /home/ubuntu/airflow/dags/goes_dag.py
[2016-12-07 21:20:14,168] {models.py:1059} INFO - Dependencies all met for <TaskInstance: goes_realtime_dag.index_file 2016-12-07 18:20:00 [queu
[2016-12-07 21:20:14,186] {models.py:1059} INFO - Dependencies all met for <TaskInstance: goes_realtime_dag.index_file 2016-12-07 18:20:00 [queu
[2016-12-07 21:20:14,186] {models.py:1248} INFO -
--------------------------------------------------------------------------------
Starting attempt 1 of 2
--------------------------------------------------------------------------------


[2016-12-07 21:20:14,198] {models.py:1271} INFO - Executing <Task(PythonOperator): index_file> on 2016-12-07 18:20:00
[2016-12-07 21:20:16,716] {connectionpool.py:805} INFO - Starting new HTTPS connection (1): signpost.opensciencedatacloud.org
[2016-12-07 21:20:16,844] {connectionpool.py:805} INFO - Starting new HTTPS connection (1): signpost.opensciencedatacloud.org
[2016-12-07 21:20:16,977] {connectionpool.py:805} INFO - Starting new HTTPS connection (1): signpost.opensciencedatacloud.org
[2016-12-07 21:20:17,091] {python_operator.py:81} INFO - Done. Returned value was: ({u'did': u'336175e9-97e8-43e1-a49e-95dfea6b45fa', u'rev': u'
```

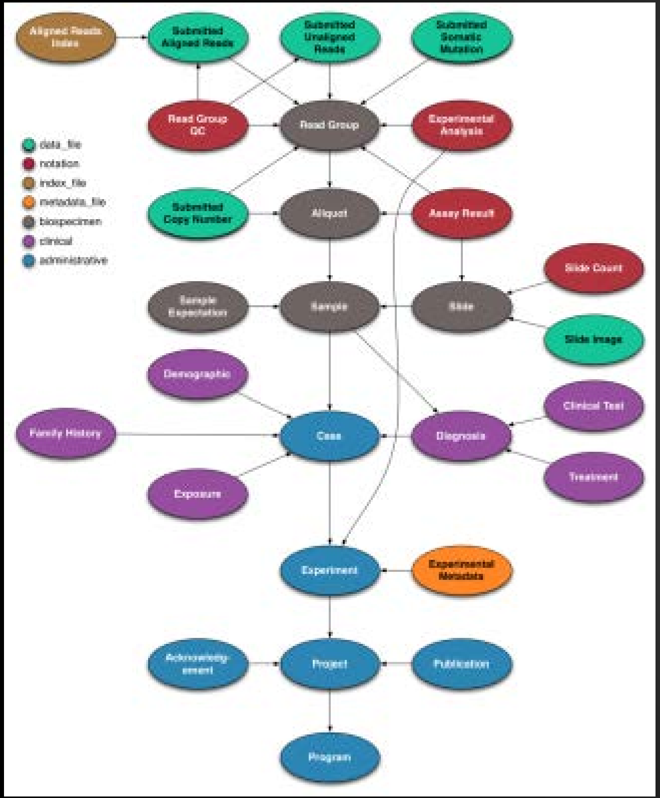
Building on Biomedical Commons

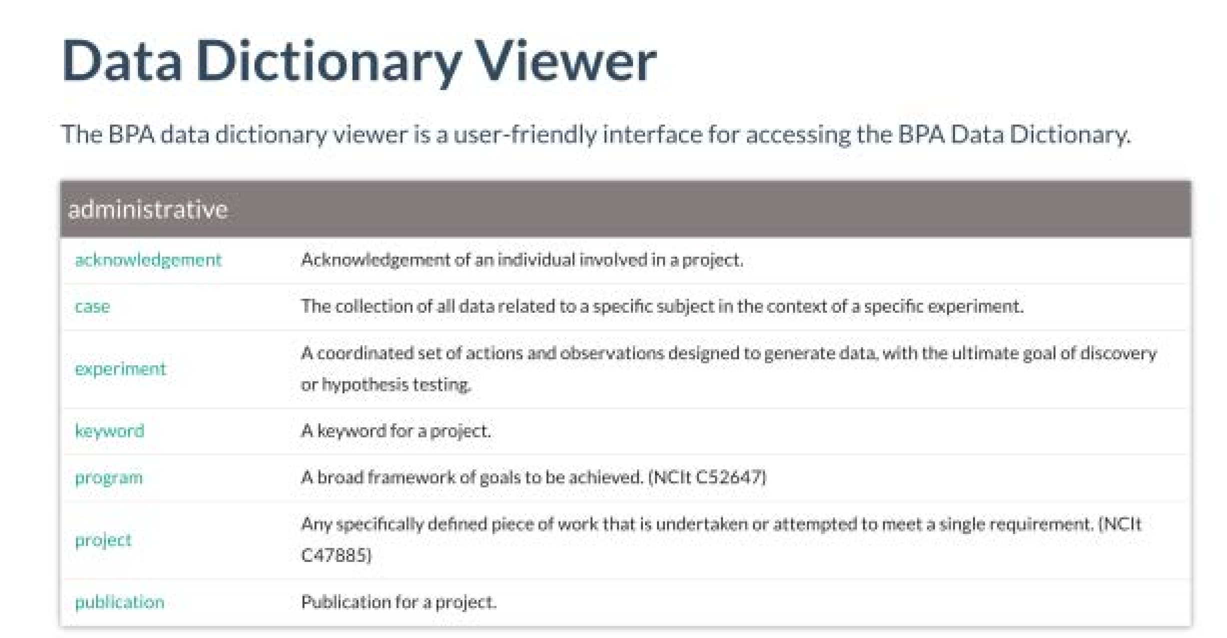
- Funded commons provide software stack improvements that we can use for the Environmental Data Commons
- NCI Genomic Data Commons
- Blood Profiling Atlas in Cancer (BloodPAC)
- Cohn Veteran's Brain Health Commons

Commons Architecture Gen 3

- Web frontend portal to backend APIs
- Submit data metadata via portal in TSV or JSON
- Query metadata using GraphQL

YAML Data Dictionary/Data Model



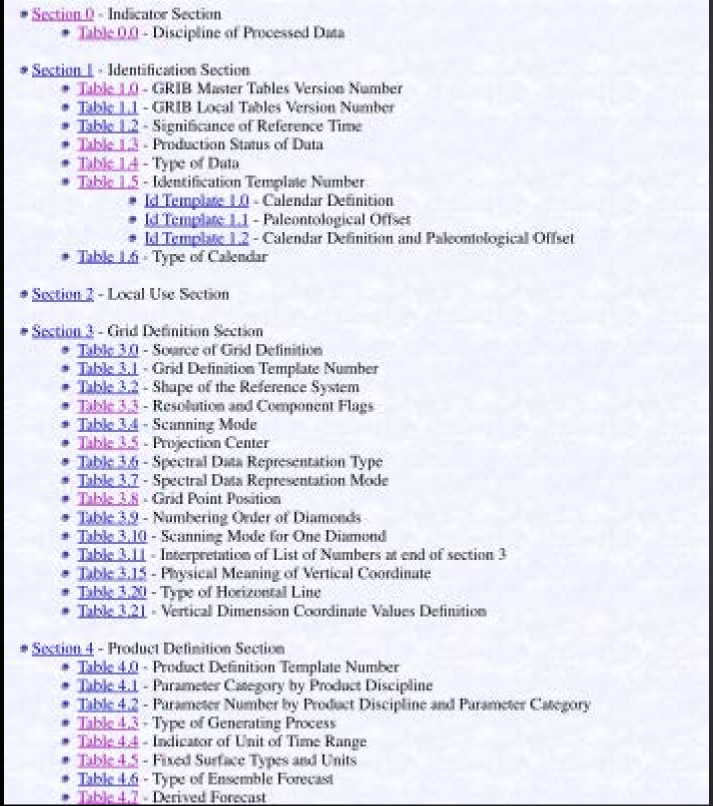


Data Dictionary Viewer

The BPA data dictionary viewer is a user-friendly interface for accessing the BPA Data Dictionary.

administrative	
acknowledgement	Acknowledgement of an individual involved in a project.
case	The collection of all data related to a specific subject in the context of a specific experiment.
experiment	A coordinated set of actions and observations designed to generate data, with the ultimate goal of discovery or hypothesis testing.
keyword	A keyword for a project.
program	A broad framework of goals to be achieved. (NCIt C52647)
project	Any specifically defined piece of work that is undertaken or attempted to meet a single requirement. (NCIt C47885)
publication	Publication for a project.

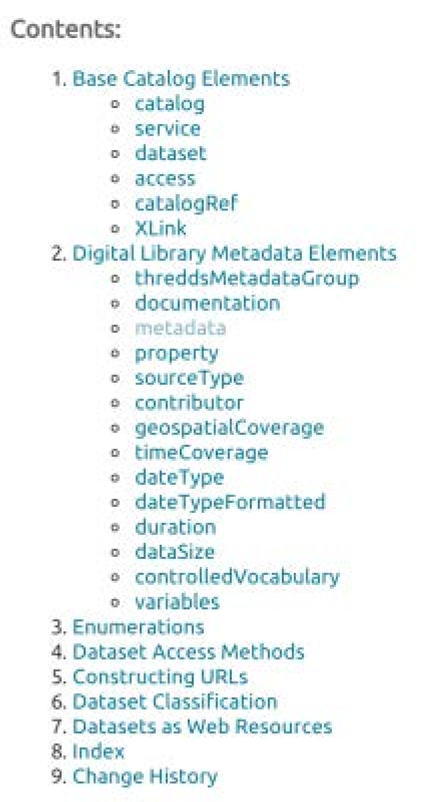
Grib Data Model



- Section 0 - Indicator Section
    - Table 0.0 - Discipline of Processed Data
- Section 1 - Identification Section
    - Table 1.0 - GRIB Master Tables Version Number
    - Table 1.1 - GRIB Local Tables Version Number
    - Table 1.2 - Significance of Reference Time
    - Table 1.3 - Production Status of Data
    - Table 1.4 - Type of Data
    - Table 1.5 - Identification Template Number
        - Id Template 1.0 - Calendar Definition
        - Id Template 1.1 - Paleontological Offset
        - Id Template 1.2 - Calendar Definition and Paleontological Offset
    - Table 1.6 - Type of Calendar
- Section 2 - Local Use Section
- Section 3 - Grid Definition Section
    - Table 3.0 - Source of Grid Definition
    - Table 3.1 - Grid Definition Template Number
    - Table 3.2 - Shape of the Reference System
    - Table 3.3 - Resolution and Component Flags
    - Table 3.4 - Scanning Mode
    - Table 3.5 - Projection Center
    - Table 3.6 - Spectral Data Representation Type
    - Table 3.7 - Spectral Data Representation Mode
    - Table 3.8 - Grid Point Position
    - Table 3.9 - Numbering Order of Diamonds
    - Table 3.10 - Scanning Mode for One Diamond
    - Table 3.11 - Interpretation of List of Numbers at end of section 3
    - Table 3.15 - Physical Meaning of Vertical Coordinate
    - Table 3.20 - Type of Horizontal Line
    - Table 3.21 - Vertical Dimension Coordinate Values Definition
- Section 4 - Product Definition Section
    - Table 4.0 - Product Definition Template Number
    - Table 4.1 - Parameter Category by Product Discipline
    - Table 4.2 - Parameter Number by Product Discipline and Parameter Category
    - Table 4.3 - Type of Generating Process
    - Table 4.4 - Indicator of Unit of Time Range
    - Table 4.5 - Fixed Surface Types and Units
    - Table 4.6 - Type of Ensemble Forecast
    - Table 4.7 - Derived Forecast

- Most useful for describing simulation outputs
- Good guide for how to define model portion of data model

THREDDS Data Model

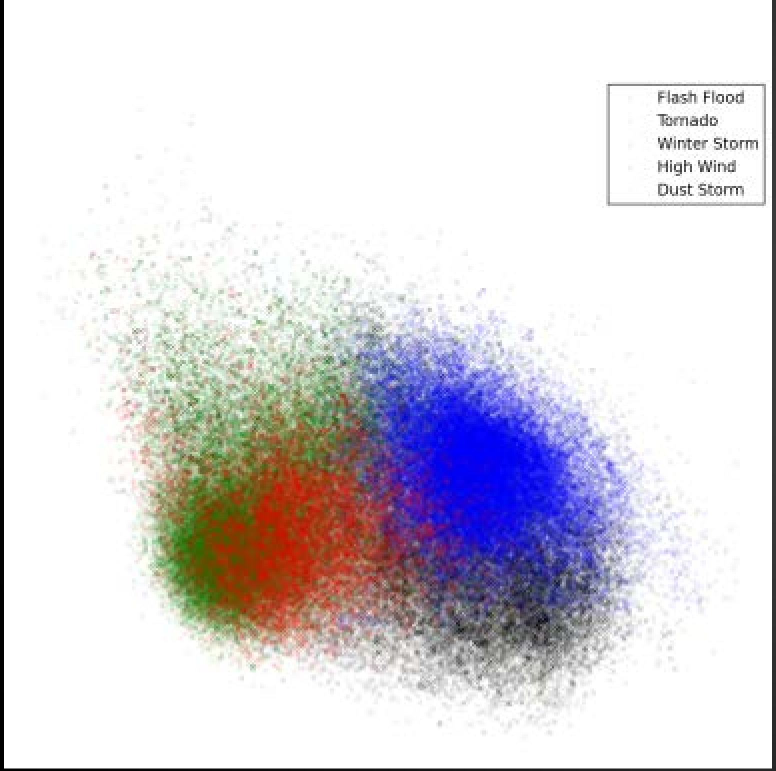


Contents:

1. Base Catalog Elements
   - catalog
   - service
   - dataset
   - access
   - catalogRef
   - XLink
2. Digital Library Metadata Elements
   - threddsMetadataGroup
   - documentation
   - metadata
   - property
   - sourceType
   - contributor
   - geospatialCoverage
   - timeCoverage
   - dateType
   - dateTypeFormatted
   - duration
   - dataSize
   - controlledVocabulary
   - variables
3. Enumerations
4. Dataset Access Methods
5. Constructing URLs
6. Dataset Classification
7. Datasets as Web Resources
8. Index
9. Change History

- XML
  - Debatable if a feature
- Good reference for what the data model should look like
- Possible to create a tool to translate between formats in the future?

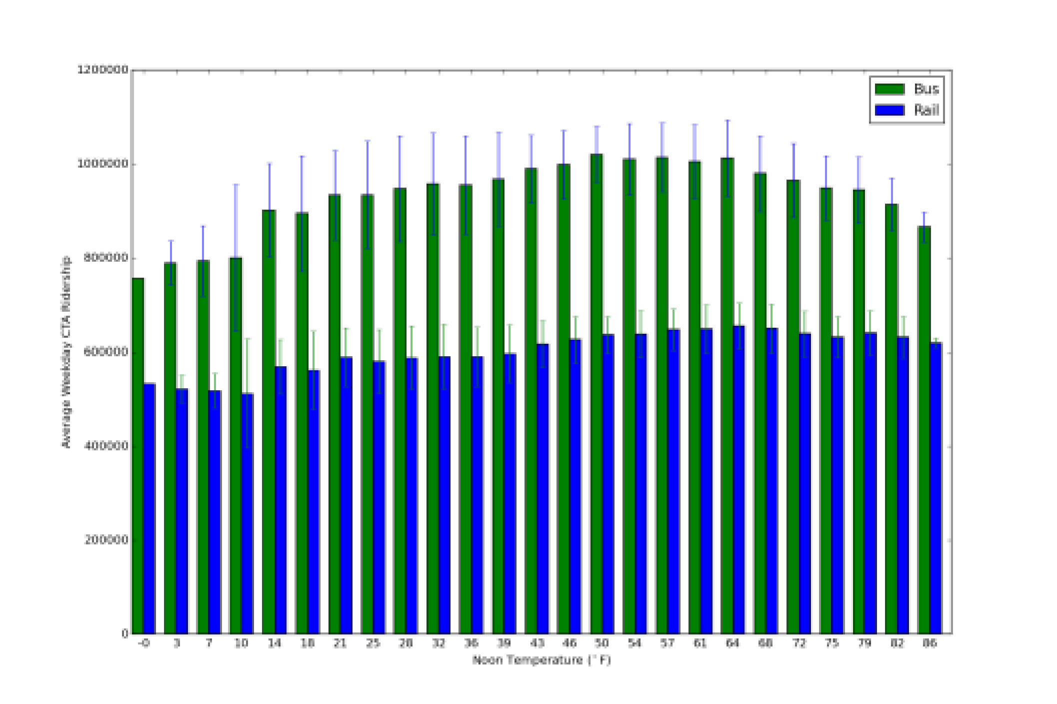
Research Examples



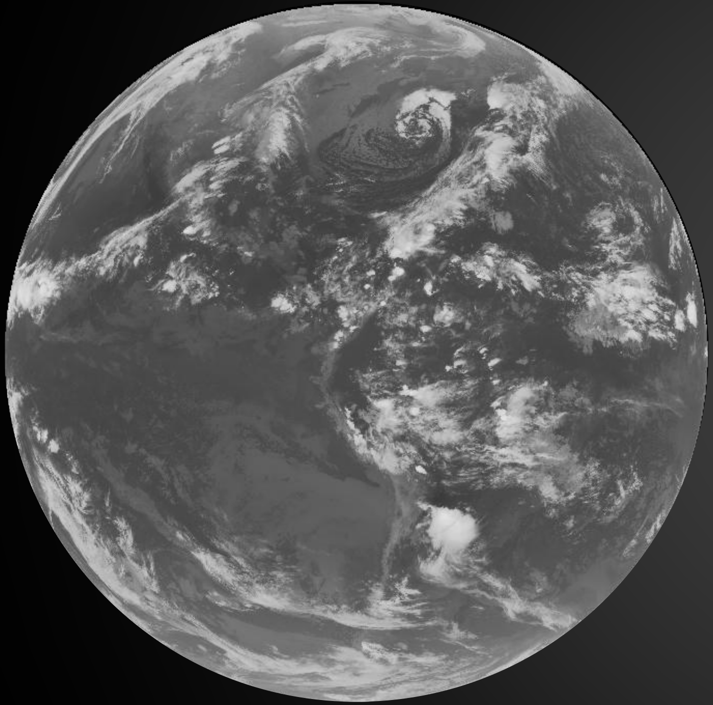
- Pull 89 variables representing county weather from CFS for each *Storm Data* event.
- Compute Principal Components Analysis to reduce dimensionality

Research Examples

- Pull Noon temperature for Chicago from CFS for daily time series of Bus & Rail rides
- Ride data from Chicago Data Portal for CTA 2001-2010



Conclusions



- Building data commons to bring data & compute together for scientific discovery
- All of the pieces are finally coming together, index services, metadata services, and workflow services
- NOAA Big Data Project facilitating easy acquisition of datasets, and dataset usage guidance
- Would not be possible to utilize as much NOAA data without the NOAA BDP

OSDC OPEN SCIENCE DATA CLOUD

http://play.opensciencedatacloud.org http://www.opensciencedatacloud.org