Topics

• Data Curation
• NOAA Big Data Program
• Data Science
• DAARWG and SAB Priorities
Data Archiving (aka Data Management)

- Data archiving focuses on the value-added activities to support the full data lifecycle: Ingest, Curation, Reuse
  - [http://www.dcc.ac.uk/sites/default/files/lifecycle_web.png](http://www.dcc.ac.uk/sites/default/files/lifecycle_web.png)

- In the satellite data world the gold standard conceptual framework is the Open Archival Information System (OAIS)
  - Submission Information Package
  - Archival Information Package
  - Dissemination Information Package
  - [https://public.ccsds.org/Publications/MagentaBooks.aspx](https://public.ccsds.org/Publications/MagentaBooks.aspx)

- Newer conceptual version is data curation should support FAIR Principles: Findable, Accessible, Interoperable, Reusable
  - [https://www.force11.org/group/fairgroup/fairprinciples](https://www.force11.org/group/fairgroup/fairprinciples)
What has this focus meant in terms of DAARWG activities:

• Early work largely focused on helping to identify areas of potential leverage where NOAA might apply its resources for maximum effect
  • Policies and procedures
  • Organizational activities
  • Some interactions on technology front

• NOAA had three long-standing data archives well-versed in data archiving leading practices

• Challenge was to encourage diffusion across the enterprise the prioritization of data as a first class object

• NOAA created a series of data management policy directives and procedures, in part due to recommendations from SAB to NOAA put forth from DAARWG

• NOAA has also kept DAARWG apprised of topics related to improving access such as the development of metadata, catalogs, and interfaces
More recently

• Along the way
  • Updated on CLASS (Comprehensive Large Array Storage System)
  • Reorganization of core data centers into NCEI
• GOES Level 1a archiving recommendation memo
• Looking to identify what other pockets of NOAA data activity are relevant
  • Social science data
• As rise of concern over replication increased, looked at what other NOAA products might be important to include as a ‘first class object’
  • Code / Models
  • Physical specimens
NOAA Big Data Project

- Started with an RFI call and followed-up with an industry day, 2014
- Announced 5, 3 year CRADAs, 2015
  - Amazon, Microsoft, IBM, Google, OCC (AMIGOs)
  - 1 year extension in the works
- Initial data put in place
  - NOAA NEXRAD data
  - Has expanded to include other NOAA data, e.g. NCDC ftp mirror
- Something of an experiment, try new collaborative partnerships, work with new technologies, contain costs, respond to demand
- Overarching goal is to improve access to NOAA data
Example BDP Success Story

- NEXRAD Radar Data: 1991- Present
- Entire NEXRAD Level 2 Archive (300 TB) was transferred from NCEI to AWS, OCC (2015-17), Microsoft, and Google
- Options: NOAA Redirects to BDP Collaborators’ services

Slide contents courtesy Dr. Ed Kearns, NOAA Chief Data Officer (June 2017)
Example BDP Success Story

- NEXRAD Level 2 Radar Data on AWS
- Ansari et al., 2017. Unlocking the potential of NEXRAD data through NOAA's Big Data Partnership
- http://journals.ametsoc.org/doi/abs/10.1175/BAMS-D-16-0021.1

Data Usage

NCEI Server Load

Increased 2.3X

Slide contents courtesy Dr. Ed Kearns, NOAA Chief Data Officer (June 2017)
BDP Partner
Links Data
Offerings

• AWS
  • https://aws.amazon.com/noaa-big-data/
• Google Cloud Platform
  • https://cloud.google.com/bigquery/public-data/
• IBM
  • https://noaa-crada.mybluemix.net/node/32
• Microsoft
  • No public service to date
• Open Commons Consortium
  • https://www.opensciencedatacloud.org/publicdata/?commons_type=Environmental

Slide contents courtesy Dr. Ed Kearns, NOAA Chief Data Officer (June 2017)
DAARWG and the BDP

• Briefings from various key leads of the Big Data Project, most recent from Ed Kearns, NOAA Chief Data Officer (June 2017)

• Issues discussed:
  • Identifying hidden costs
  • Challenges of potential multiple copies of data
  • Tracking usage and receiving credit
  • Storing data in the cloud versus curation of data in the cloud
  • Privatizing a public good

• DAARWG has been enthusiastic about the experiment
Data Science

- Is there a common understanding of what Data Science is?
  - Analytics (statistics + big data)?
  - Science of data?
  - Techniques, tools, standards, knowledge, communities of practice around making data reusable and interoperable?
DAARWG and SAB Topics

• Finalize revisions to DAARWG Terms of Reference
• Specialized curation requirements for:
  • Information products
  • Decision-support data
  • Social science data
  • Models and software
  • ‘Omic’ data
  • Other data platforms, e.g. IoT, drones
  • Citizen science
• Explore information sharing with other working groups