Long-term Preservation and Stewardship Breakout

Wednesday, May 26, 1:00 to 3:30 PM, SSMC-3, Room 4527

Session Leads: Steve Del Greco and Nancy Ritchey

Overview: This breakout session focuses on issues related to the long-term preservation and stewardship of NOAA's environmental data.

Objectives

- Familiarize participants with the "NOAA Procedure for Scientific Records Appraisal and Archive Approval" process
- Outline the way forward for the development of a "How to Archive" concept of operations building on the work of the Archive Architecture Team
- Discuss the Open Archival Information System - Reference Model (OAIS-RM) as a common language for discussing the long-term preservation and stewardship of information
- Discuss preservation and stewardship standards such as configuration management, data migration, and common data model
- Data stewardship and information preservation in the “cloud;” federation, dynamic processing, and work flows in a widely distributed system

  (above taken from Registration packet)

- Implementation
  - Referred to NOAA’s What to Archive Procedure flow.
  - Begins with “Data Center receives request to archive”
  - Ends with: “Ingest and archive data”
- Appraisal process for retiring a data set/product; 30-day waiting period for public comments.
  - This process should apply to both the removal of the data product as well as the removal of the original data - any time you’re removing it from public consumption; including versioning.
- NCDC creating a tool called TRACK. Hard to figure out which data you’re in the midst of working on, so this helps handle all the requests.

Issues

- Resources
  - Staff
  - Appraisal team – you need experts on the team to give their opinions on the data
  - Time – informal appraisals are usually pretty quick. Formal (i.e. GOES-R) has been going on for over a year.
  - Funding – Data Management plan really need to include provisions for this
- Looking at the possibility of doing data request by request, and only processing the level 0 and level 1 data (not the level 2 products).
- A lot of interaction involved with an in-situ data set, and can create complicated stewardship issues.
- Moving forward, you’re going to need human resources to manage the stewardship.

- What is the incentive in academia and other outside resources to help appraise data sets?
    - Roundtable staff serves as a data set appraisal team; everyone serves time on it, assesses it, then does a yes/no vote on it.
    - Or “friends of a data center” pull in folks who are retired, who have a lot of experience and some spare time, in a rotational way.
    - We must create some kind of incentive for these “friends” to take part in the appraisal of data sets.
- The research community is including links in their papers, so the expectation is increasing that when you read a paper, you’ll be provided with a link to the data that was used to generate that paper.
    - Are we an archive of data or are we an archive of algorithms?
    - It’s a changing way of doing business.
- Take a snapshot of the computing system so that you can recreate the hardware and software to recreate the data product.
- You need experts on your appraisal team to know what data is restricted.
    - Does the public have access to that raw data because it was bought by the government?
    - Much of the data that’s acquired by the NSF is used to regulate industry, so there can be a lot of interest in it.
- If you don’t have the human resources to archive it the way you want to, then perhaps it just goes into the “deep archive.” Although some dislike that term.
- Who, how, where, how to pay for it? When the money runs out for some outside data producer, they may look to a national data center.
    - What if we want to steward the data through academia, and they don’t have the funding to do it?
- How do we determine who amongst us will serve as the data center that becomes the home for the “homeless data” that’s generated by outside agencies but is no longer being stewarded by these agencies?
- What is the scope of what each one of the data centers are to do?
    - Is it a library function?
    - Provide additional documentation? Wider still?
- Analogy: If you have a nuclear power system, you have to deal with the nuclear waste.
- Space and Earth-based platformed data coming in at the rate of terabytes/year will increase over time to petabytes/year.
- OAIS RM is an ISO Standard model for information preservation.

Plan for a NOAA Archive Concept of Operations:

- Sophistication aspect of the preservation is one way to look at the problem.
- Current attitude is that if we get the data, we need to hire someone who’s an expert on the data to steward it? However, we need to look at alternative ways of getting that done.
- Comment: Throw everything in the Library of Congress?
- Data Centers were set up for preservation and stewardship of data.
- On the OAIS-RM, the designated community is fundamental to an archive.
    - The designated community or the steward will use the parts of the system that they need once a dataset has been identified as important.
    - Changes your cost model because you may have to pay for the stewardship of a dataset for many years before it becomes valuable to its designated community.

- Levels of service idea and tiered support is an idea that needs to be explored.
- Has there ever been any thought to making one of the PIs
  - Model is capture the result while they have the funding, that is:
    - Lifetime of the grant (attach a plan to each grant)
    - Lifetime of their personal computer
  - Associated cost for the long view?
    - PIs are encouraged to include the cost of data management in their funding request.
    - Transfer of funding to the data center activity performing this task.

Federating of Databases

- Cloud definition: a computing storage system that we can tap into when we need it.
- Tiered data services
  - Tiers within tiers
  - Tier 0: long term storage
- What is THE problem?
  - Data volume? No
  - Network bandwidth? No
- Problem is Scientific Data Stewardship
  - This is the long term problem we can't just throw money at because we still have to figure out just how to do it.
- What's a Federation?
  - You need to move the data around
  - Geodesic dome analogy… the more nodes, the stronger it is. You can't build a dome with 2 nodes, you need more – and the more the better.
- Example of Federation: NDBC > Data center > NOMADS > CLASS > (back to NDBC)
  - Bring in ESG (Earth System Grid) and insert them into the existing federation.
  - Bring in cloud resources
- Finding a sweet spot, we're in this "brave new world" where a lot of things are possible where many applications haven' been born yet. However, it also opens us up to a lot of risk, when you rely on too many interdependencies.
  - Some cringe at the notion of deep archive, because archive should mean access, even if it's by a smaller group of users in a tailored way.
- Is there a way they can retain the long term preservation as catastrophic loss insurance?
- CLASS is designed to serve Data Centers and Centers of Data.
- Even if the data is someplace else, another group can be the stewardship of the data. You can be the steward of a data set by virtue of being one of the chief stakeholders. Who the stake holders are is more important than proximity to the data.
- Many of the same challenges that are being dealt with in the health care industry with regard to personal health care related information are being dealt with now. There's the potential for cross-pollination of stewardship ideas between the two industries.
  - There's either a financial incentive or business incentive to participate in a federated model.
  - Another industry that deals with such data stewardship issues is the banking industry.
  - Although, it's all coming down to digital data management, which transcends every industry and comes into our homes.

    - In the case of a smaller center of data, if they have a simple service provided to interact with CLASS to help with preservation. That might be incentive enough for them to participate.
- When the amount of data is small, doing regular backups is quick and simple. For example, Fisheries might be able to backup all their data on site.
    - In such a case, the local outfit may not want the service of data backup, except to co-locate metadata services - making their metadata available to the cloud.
    - So, if data backup is not something they need from a data center, then what kind of services would Fisheries like?
- Broadcasting is another industry confronted with a monumental data management task, as video is the largest kind of data by a couple orders of magnitude.
- Who is going to tell the smaller data center, “You can’t run a data archive out of your facility anymore?”
- Being a trusted source has got to be part of your “business model”
    - Digital signatures have to be used to guarantee integrity of data
    - Banking industry is the example for data security.

Some summary points:

- OIAS -RM used as a reference model.
- Need for an archive ConOps needs to get spun back up - needs to be fully vetted.
- Preservation and Planning needs to be moved to the top of ConOps.
- Fitting into the new landscape, fitting into the cloud and a federation of data.
- Volume across stewardship.
- Federation should help ameliorate problems once implemented across NOAA.
- Sweet spot for processing in house; when to push data to a high performance data center .
- SOA and rules based data management.
- Looking at other industries to see how they federate their databases and work in the cloud: Banking, Health care, and Broadcasting.