NCEI PASSIVE ACOUSTIC DATA WORKSHOP

A two-day workshop was held at NCEI in Boulder, CO on June 10-11, 2015 for NCEI, NOAA Fisheries Science Center, Office of Science & Technology, Bureau of Ocean Energy Management, and Navy staff. The workshop brought attendees together to outline the progress of the passive acoustic pilot project and discuss the future direction of long-term passive acoustic data stewardship.

June 10, 2015

DATA CENTER CONSOLIDATION AND ARCHIVING

To provide a more seamless exterior for data providers and users to submit and access archived data, the National Geophysical Data Center, the National Oceanographic Data Center, and the National Climatic Data Center have merged to form the National Centers for Environmental Information (NCEI). ***The transition will not happen overnight but regardless of the institution's name,*** Carrie Wall ***(project lead) and*** Chuck Anderson ***(data manager) remain the points of contact for passive acoustic data stewardship.***

There are many data management policies and directives driving the current push to ensure publically funded data are archived and made accessible to the public. The biggest driver is the White House OSTP memorandum increasing access to the results of federally funded research (Providing Access to Research Results – PARR) and the NOAA and NSF plans in response to PARR. ***See the hyperlinks in Sue's presentation (see McLean_NCEI-Overview-June 2015-with PARR) for detailed information on PARR and other directives as well as the PARR presentation provided later by Mark Brady, NMFS' Information Architect (see Brady_PARR Briefing March 2015).***

Under NCEI's Tiers of Data Stewardship, the pilot is nearing Tier 1: Long Term Preservation and Basic Access. Several steps will need to be taken to ensure Tier 1 requirements are met and to move to Tier 2: Enhanced Access and Basic Quality Assurance. ***The tiers represent a scale of service from which cost models are based.*** The more basic the requested services, the more cost effective. Further, the more complete and standardized the metadata and raw data are, the more cost effective archiving will be.

STATUS OF THE PILOT PROJECT

The progress of the pilot project initiated in July 2014 was presented. To date,

- a database schema has been developed off the deployment fields within Tethys
- an ingest pipeline has been established, which consists of
    - moving files off submitted drives
    - FLACcing the files (i.e., lossless compression)
    - aggregating files into 4 hour windows
    - tarballing the file aggregates
    - and moving them to the archive

These steps were tested on a portion of the 3TB of data submitted by NEFSC. ***A few additional improvements are needed but in general the process appears to work. We await Catherine Berchok's***

data submission to further test the robustness of the system. We have not developed a prototype for a map viewer yet but will look into that in the coming months.

Challenges identified throughout the pilot project include

- NCEI-CO is not readily capable of handling data submission provided on internal drives
- metadata exported from Tethys are not ISO compliant and will need to be expanded
- there is a limit on the number of channels for a file to be FLACced, format dependent
- multi-channel files representing separate geographic locations is a challenge, particularly for properly capturing the metadata
    - Future work may involve a mechanism where NCEI unstitches such files, archives them as separate deployments, and re-stitches the files upon request

BEYOND THE PILOT

In order to transition the pilot project into a fully operational archive, we will need to

- secure funding
    - Potential sources are NMFS S&T, BOEM, NSF, and possibly the Navy
- extend the Tethys exported metadata into ISO compliance
    - Several fields need to be added to Tethys (e.g., information gain, calibration, duty cycle, anti-aliasing filtering). We will follow up with Marie to discuss these items
    - A data packaging tool, similar to what is used for the water column sonar data archive, may be developed to assist data providers in providing additional metadata information (e.g., contact information, title, project description), and to easily conduct an MD5 checksum, a data integrity check during file transfer
    - Align metadata fields with ASA standards and IOOS' metadata convention
- establish the syntax for assigning Digital Object Identifiers (DOIs) to the datasets
    - Something similar to the water column sonar data archive DOIs
        - Alaska Fisheries Science Center (2012): Water Column Sonar Data Collection (DY1207, ME70). National Geophysical Data Center, NOAA. doi:10.7289/V5KK98PX [access date]
- identify the next dataset to archive
    - The first recovery from the Ocean Noise Reference Station Network project may be a good start but the group/SOST will create a prioritized list to guide our archive efforts
- and develop a map viewer for the discovery and access of these data
    - It will be similar to the water column sonar data map viewer
    - Filter functions to include
        - Damaged data
        - Spatio-temporal constraints including specific time windows
        - Recording type (continuous vs. intermittent and single vs multi-channel)
        - Platform type (fixed, glider, towed, drifter)
        - Frequency range
        - Instrument type
        - Sensor depth
        - Organization/Funding source

The above steps meet the requirements of PARR.

GETTING TO KNOW YOUR DATA

I summarized the volumes of backlogged and annually collected passive acoustic data based on the information you provided during your presentations (see PAD volume estimates). ***A conservative total for backlogged data is 1093 TB with 277 TB collected annually!*** Dan Kowal, head of the Standards and Evaluation Division, provided the following cost estimates for archiving data for 10 years:

- Large volume (200 TB) = $137,000
- Small volume (24 TB) = $40,000

Briefly ignoring the massive amounts of backlogged data, archive costs for currently collected data are likely ~$200k/yr. Additional costs will need to be considered and the above estimates may be adjusted as NCEI further define its cost models.

I followed up with Mark Brady regarding Shannon's question about R&D type data: ***If these data are not going to be used/referenced in the future to inform management decisions, then they can be exempt from PARR's requirements.*** Essentially, if these data represent scratch paper they fall outside of PARR and you can use your own discretion to determine what data sets should/should not be archived.

SUMMARY AND NEXT STEPS

The level of interagency collaboration became clear throughout the workshop. We (NCEI) will continue discussions with all groups involved. ***NCEI's passive acoustic data archive will focus on the long-term preservation of raw acoustic data; these efforts will be augmented by linking to processed acoustic data web portals (e.g., OBIS-Seamap).***

Further the enormous volume of acoustic data within NMFS and collaborating agencies highlights the need to expand the archive in increments, as funding allows. ***Mark Brady is aware of the large volume of these data and obtaining waivers to postpone looming PARR deadlines should be explored.***

Interest was expressed for holding these workshops annually. ***The late February/March timeframe was suggested to coordinate with the annual water column sonar data workshops.*** The group had varying availability during this time; a doodle poll will be used to coordinate schedules.

Interest was also expressed for NCEI-CO to host an internal Google Site for the passive acoustic data project. This will be developed in the coming weeks.