Overview of Deepwater Horizon Long Term Data Management

DWH long-term data management workshop
June 7-8, 2017
Lauren Showalter
Program Officer – Information Science
National Academies of Science Gulf Research Program

Goals:
• To ensure the legacy of the science from the DWH disaster is accessible to researchers studying future disasters in the region and around the world
• To improve the quality of science coming out of the funds from the DWH disaster and ensure the research products are well documented and in stable formats
Data Archives

In the field of data management, the terms "archive" and "repository" often are used interchangeably. Within the Federal government, however, the term "archive" is specific to the mission and activities of the National Archives and Records Administration (NARA). Only NARA, or a Federal entity officially delegated by NARA for the long-term curation of specific products, should be referred to as an "archive."

From Open Archival Information System (OAIS)

...an archive, consisting of an organization, which may be part of a larger organization, of people and systems that has accepted the responsibility to preserve information and make it available for a designated community. It meets a set of such responsibilities as defined in this International Standard, and this allows an OAIS archive to be distinguished from other uses of the term "archive". (from iso.org)

From NOAA and U.S. National Archives and Records Administration (NARA)

The NOAA National Data Centers are tasked with storing environmental data and making this data available to researchers, scientists, and anyone else that has a need for it, as well as in support of NOAA’s mission. Destroy/delete 75 years after cutoff upon approval by NOAA and NESDIS stakeholders. A longer retention may be necessary for research purposes.
How do we know we are restoring to previous conditions?
  • What information is available to know that
  • Comparable metrics and baselines for monitoring and restoration activities
  • What is the new baseline for the GoM since DWH?
  • Want to be able to look back at DWH data to answer questions for future spills
Documentation of data is essential to ensure that future users understand how the data was collected and who to contact with questions.

- ISO 19115 standard should be adopted as much as possible:
  - This is what the federal government is using
  - Other standards should be able to be transformed into ISO

Darwin Core could be considered for biological data, for specific repositories.

**Metadata Creation**

ISO 19115-2 Metadata Editor

This section collects contact information about the person primarily responsible for the creation and maintenance of the dataset. Provides future researchers the means to contact the dataset creator if additional information is needed.

- **Name**:
  - The name of the individual responsible for the creation of the authority of the dataset.

- **Organization**:
  - Responsible individual’s organization, generally this is the lead agency and Department where the person resides. This information can be found at research.gov.

- **Position**:
  - Position of the responsible individual within the organization, e.g., Assistant Professor.
XML Metadata

Metadata Training
Data Standards

- Identification of standards early in the process and get community buy in
- Standards need to be adequately communicated to data collectors

Federal Mandates

- The Digital Government Strategy and Open Data Policy were developed for the Government to better deliver information (data) and services.
- Federal agencies are under certain mandates that could inhibit data from being accepted if not properly formatted and documented
- *Common Framework for Earth-Observation Data*, March 2016, Office of Science and Technology Policy

- These standards need to be properly communicated to the data collectors
Data Sharing

- Length of time from collection to sharing
  - Real time data – data that is shared as soon as it is collected
    - Cruise data
    - Satellite data
    - Buoy data
  - Other data is shared depending on:
    - Funder/publisher requirements
    - Federal or state mandates
    - Requirements of other collaborators (foreign, private, industry, etc.)

Data Holds

- If data is to be held for any reason the documentation of that data should begin before it is submitted for public access.
- Groups that start documentation before the data is collected have a leg up when the data is ready for publication
- Tracking of data from project onset is essential
Data Citation

- support proper attribution and credit
- support collaboration and reuse of data
- enable reproducibility of findings
- foster faster and more efficient research progress
- provide the means to share data with future researchers

Persistent Identifiers

- A Digital Object Identifier (DOI) is a commonly used type of identifier that is used to link to digital objects.
- Use of a persistent identifier makes data search and accessibility easier for future users
- Open Researcher and Contributor ID (ORCHID) is a persistent identifier for researchers
Providing services that allow users to retrieve data for exploration, analysis, or decision making

Rely on sets of common standards and protocol (e.g. OPeNDAP, WMS, WCS, ERRDAP, FTP, SOS)

Often community-driven

Need for both human access and machine-to-machine access
Interoperability

- Technology methods (examples):
  - OpenDAP
  - THREDDS
- Machine to machine data tools
- This allows for better and easier data synthesis

Distributed Data Frameworks
Good data management practices allow data to be easily reused and synthesized to develop useful products.

Display and manipulation of integrated data:
- ERDAPP
- Cesium
- ESRI
- And many more
Telemetry Data

Visualization examples