VIRTUAL COLLECTIONS:
An Earth Science Data Curation Service

Kaylin Bugbee¹ (kbbugbee@itsc.uah.edu), Rahul Ramachandran², Manil Maskey², Patrick Gatlι²
¹The University of Alabama in Huntsville, ²NASA MSFC

The role of Earth science data owners has traditionally been to maintain central archives that serve them as a primary storage site and information resource. Earth science data centers must now move beyond simply serving as an archive to offering innovative data services to user communities. A virtual collection, the end product of a curation activity that searches, selects, and synthesizes Earth science data and information resources around a specific topic or event, is a data curation service that improves the discoverability, accessibility and usability of Earth science data and also supports the needs of numerous users. Virtual collections will enable the generation of local and national datasets through the discovery of relevant data and metadata. A virtual collection will provide a framework for understanding the complexity of Earth science data and will highlight a virtual collection case study created at the Global Hydrology Resource Center data center.

https://ghrc.nsstc.nasa.gov/home

CURATION

The steps to creating a virtual collection follow the governance model of searching, selecting and synthesizing Earth science data, metadata and information into a cohesive and useful collection (Ramachandran et al., 2016).

1. Establish pre-curation framework by defining the goal, the audience, and the fitness criteria of the virtual collection.
2. Search for all related data and information for possible inclusion in the collection.
3. Select relevant data using defined fitness metrics and temporal/spatial bounds.
4. Synthesize data and information. Synthesis occurs on a spectrum from data accessibility to data usability and is based on the desired goal of the virtual collection.

LESSONS LEARNED

Pre-curation framework:
- Identifying interesting events and the related framing criteria is time consuming. For the GCPEX case study, mission reports, campaign blogs and peer reviewed publications were surveyed to identify an interesting event. Identifying and documenting relevant events is a critical step to streamline the search process.
- The search for data was limited to data provided by the GHRC to simplify the case study. Limiting data to the GHRC made the search process relatively simple because GHRC has access to several Earth science datasets.
- Incorporate spatial information. Accurate spatial information was only introduced at the collection level. While the dataset level metadata did include spatial information, this spatial information was not provided at the collection level. Due to these constraints, the GCPEX uses required that grantees be submitted to the requested spatial boundaries using OPeNDAP.
- Synthesis:
  - Data bundling: There are many ways to bundle the submitted grantees including individually listing the submitted grantees, by providing zip files, or by providing a single zip file. Selecting one method is totally dependent on the downstream applications that use the bundle.
  - Data formats: For dynamic data subsets of the grantees using OPeNDAP, the metadata should include standard formats such as netCDF or HDF. However, proprietary ASCII or XLS formats were encountered for grantees across several datasets. Thus, it was necessary to develop a format translation utility to convert non-standard grantees into netCDF format. It should also be noted that OPeNDAP-based subsets of the data is only possible if there is a grid type defined for the parameter that was needed for subsetting.