VIRTUAL COLLECTIONS: An Earth Science Data Curation Service

Kaylin Bugbee¹ (kbugbee@itsc.uah.edu), Rahul Ramachandran², Manil Maskey², Patrick Gatin²
¹The University of Alabama in Huntsville, ²NASA MSFC

1. Establish Precuration Framework
   - Identify Data Center Goals/Access.
   - Define Fitness Criteria for the Virtual Collection (Conduct a Scoping Questionnaire)
   - Select Audience

2. Search for Data and Information
   - Conducted by Domain Experts
   - Conducted using fitness criteria, spatial and temporal bounds

3. Select Relevant Data and Information
   - Conducted using fitness criteria, spatial and temporal bounds

4. Synthesis
   - Conducted

USE CASE

- The Global Hydrology Resource Center (GHRC), one of NASA's twelve Distributed Active Archive Centers (DAACs) curated a virtual collection that highlighted a scientific process relevant to the DAAC's data holdings and that addressed a need for data bundles from the user community.
- A use case was identified from NASA's GPM Dual-Sensor Precipitation Experiment (GCPEx) field campaign data that is archived at the GHRC. The GCPEx field campaign was undertaken to collect a 3-D high-resolution dataset of snowfall physical properties and radiative properties that will be utilized to develop snowfall retrieval algorithm for the GPM.

Pre-curation framework:
- Goal: Data usability
- Audience: Domain experts or initial user community
- Topic: Snow microphysicals
- Fitness criteria/selection criteria: What is the 3-D structure of falling snow and how does its variability affect remotely sensed retrievals?

Search:
- A search was conducted for datasets related to:
  - Microwave remote sensing on both polarized and airborne platforms.
  - Snow particle size and snow-water equivalent measurements also from both ground-based and airborne platforms.

Select:
- Data was curated by:
  - Confirming data was collected during the identified temporal period of February 24, 2012.
  - Selecting data with parameters most relevant to the fitness criteria/selection question.
  - Some data was pre-culled from being selected due to data formatting challenges.

Synthesis:
- Identified data was synthesized with the goal of increasing data usability. The GCPEx virtual collection includes:
  - Metadata container – Metadata describing the virtual collection was created in collaboration with a domain-expert.
  - Data bundle – Python notebook for subsetting and selection by parameter (https://github.com/producedata/research/metadata/GCPEx/Golden-Class-bundle)
  - Contextual Documentation – A micro article was written for the virtual collection. Micro articles are short, interesting documents that bring together data and key science concepts. Micro articles create a knowledge base for users by curating knowledge around the science thematic areas of the data center and the data offered by the data center. Micro articles are curated by both Earth and data scientists to ensure the accuracy and trustworthiness of the provided information. The microarticle for the GCPEx virtual collection describes the February 24, 2012 event and the science phenomena. The micro-article also provides information on the member datasets that were subsetted within the virtual collection.
  - Finally, relevant publications and reference sources are also listed.

Finally, the GCPEx virtual collection was also published using the normal GHRC publication work flow. As a result of this precuration effort, a DOI was created for the collection (https://doi.org/10.5067/GCPEx/S/MULTIPLEDATA101).

LESSONS LEARNED

- Pre-curation framework:
  - Identify interesting events and the related framing criteria is time consuming. For the SCPE use case, mission reports, campaign blogs, and peer reviewed publications were surveyed to identify an interesting event. Identifying and documenting relevant events is a field campaign takes place days to sometimes weeks prior to occurrence.

- Search:
  - The search for data was limited to data provided by the GHRC to simplify the case. Limiting data to the GHRC made the search/step relatively simple because GHRC provides collection around each field campaign. Searching for data outside the data center requires more time and effort.

- Select:
  - Metadata quality is important and was a limiting factor in selecting data for the virtual collection. Information gaps are in the metadata included:
    - Incomplete temporal information: Temporal information was provided only at the collection level and in most cases was nonexistent at the granule level. This limitation meant that granules had to be manually extended.
    - Incomplete spatial information: Sometimes spatial information was only provided at the collection level. While the granule level metadata did include spatial information, this spatial information matched the collection level coordinates. Due to these constraints, the SCPE use case required that granules be subsetted to the required spatial collection bounds using OPeNDAP.

- Synthesis:
  - Data bundling: There are many ways to bundle the subsetted granules including individually testing the subsetted granules, by providing zip file or tar bundles, etc. Selecting one method is totally dependent on the downstream applications that use the bundle.
  - Data format: For dynamic subsetting of the granules using OPeNDAP, the granules needed to be provided in standard formats such as netCDF or HDF. However, proprietary ASCII or XLS formats were encountered for granules across several datasets. Thus, it was necessary to develop a format translation layer to convert non-standard granules 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 subsetted.