VIRTUAL COLLECTIONS: An Earth Science Data Curation Service

Kaylin Bugbee¹ (kbbugbee@itsc.uah.edu), Rahul Ramachandran², Manil Maskey², Patrick Gatlin²
¹The University of Alabama in Huntsville, ²NASA MSFC

1. ESTABLISH PRECURATION FRAMEWORK
2. SEARCH FOR DATA AND INFORMATION
3. SELECT RELEVANT DATA AND INFORMATION
4. SYNTHESIS

USE CASE

- The Global Hydrology Resource Center (GHRC), one of NASA's twelve Distributed Active Archive Centers (DAACs), curates a virtual collection that highlights a scientific process relevant to the DAAC's data holdings and that also addresses a need for data bundles from the user community.

Pre-curation framework:
- Quality: Data usability
- Audience: Domain experts or initial user community
- Topic: Snow microphysics
- Fitness criteria/stakeholder: 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 selected by:
  - Confirming data was collected during the identified temporal period of February 24, 2012.
  - Selecting data with parameters most relevant to the fitness criteria/satellite question.
  - Some data was pre-processed from being selected due to data formatting challenges.

Conducted by domain expert.

Cull results using fitness criteria, spatial and temporal bounds.

Synthesis:
- Identified data was synthesized with the goal of increasing data usability. The GCPEX virtual collection includes:
  - Metadata container – Metadata describing the virtual collection created in collaboration with a domain expert.
  - Contextual documentation – A microarticle was written for the virtual collection. Microarticles are short, interesting documents that bring together data and key science concepts. Microarticles create a knowledge base for users by curating knowledge around a science thematic area of a data center and the data offered by the data center. Microarticles 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 microarticle also provides information on the member datasets that were sub-setted within the virtual collection. Finally, relevant publications and reference sources are also listed. (https://ghrc.msfc.nasa.gov/home/micro-articles/microarticleon-event-during-gcpx-field-campaign/)

- Finally, the GCPEX virtual collection was also published using the normal GHRC publication workflow. As a result of this publication effort, a DOI was created for the collection (http://doi.org/10.5067/GCPEX5S/MULTIPLEDATA/141).

The steps to curating a virtual collection follows the general process of searching, selecting and synthesizing data science data, metadata and information into a cohesive and useful collection (Ramachandran et al., 2016).

Pre-curation framework:
- Identify interesting events and the related framing criteria is time consuming. For the SCPEX use case, mission reports, campaign blogs and peer reviewed publications were surveyed to identify an interesting event and documenting relevant events as a field campaign takes place aids in streamlining the search process.

Search:
- The search for data was limited to data provided by the GHRC to simplify the use case. Limiting data to the GHRC made the search step relatively simple because GHRC creates collections 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 in the metadata included:
  - Incomplete temporal information
  - Temporal information was provided only at the collection level and not more specific at the granule level.
  - Incomplete spatial information
  - Spatial information was only provided at the collection level. While this granule level metadata did include spatial information, the spatial information maintained the collection level coordinates. Due to these constraints the GCPEX use case required that granules be subsetted to the requested spatial bounds using OpenDAIP.

Synthesis:
- Data bundling: There are many ways to bundle the submitted granules including individually listing the submitted granules, by providing zip files or tarball 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 OpenDAIP, the granules needed to be provided in standard formats such as netCDF or HDF. However, proprietary ASCI or XLS formats were encountered for granules across several datasets. Thus, it was necessary to develop a formatting tool that could convert non-standard granules into netCDF format. This tool should also be noted that OpenDAIP-based subsets of the data is only possible if there is a grid type defined for the parameter that was needed for subsetting.

System limitations: The Python notebook could not be directly hosted within the GHRC data center due to security concerns. Therefore, only a static or already executed version of the notebook is being hosted. This limitation decreases accessibility to the collection and in essence makes the collection non-virtual.