VIRTUAL COLLECTIONS: An Earth Science Data Curation Service

Kaylin Bugbee¹ (kbugbee@itsc.uah.edu), Rahul Ramachandran², Manil Maskey³, Patrick Gatlin³

¹The University of Alabama in Huntsville, ²NASA MSFC

Establish Precuration Framework
Search for Data and Information
Select Relevant Data and Information
Synthesis

Conducted by Domain Experts
Cull Results Using Fitness Criteria, Spatial and Temporal Bounds

ACCESSIBILITY
Metadata Annotation (Keyword Tags)

USABILITY
Metadata Container
Data Bundle
Contextual Documentation

USE CASE

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

A use case was identified from NASA's GPM-DARWIN 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 database of snowfall physical properties and radiative properties that will be utilized to develop snowfall retrieval algorithms for the GPM.

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

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

Select:
- Datasets were culled by:
  - Confirming data was collected during the identified temporal period of February 24, 2012
  - Selecting data with parameters most relevant to the fitness criteria/science 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
  - 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 basis for users by curating knowledge around the science thematic areas of a 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 microarticle 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 publication effort, a DOI was created for the collection (http://doi.org/10.5067/GCPExCS/MULTIPLEDATA101)

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

Curation

The steps to curating a virtual collection follows the governance model of searching, selecting and synthesizing Earth science data, metadata and information into a single, 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 potential inclusion in the collection.
3. Select relevant data using defined fitness criteria and spatiotemporal 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 use case, mission reports, campaign blogs, and peer reviewed publications were surveyed to identify an interesting event and identifying and documenting relevant events is a field campaign's time- and resource-intensive 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 supports 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 in the collection level and in most cases was not present at the granularity level of the data set. This limited the granularity level metadata for the collection.
  - Incomplete spatial information: Accurate spatial information was only provided at the collection level. While the granularity level metadata did include spatial information, this spatial information matched the collection level coordinates. Due to these circumstances, the GCPEx use case required that granules be subsetted to the required spatial subset bounds using OffNDAP.

Synthesis:
- Data bundling: There are many ways to bundle the subsetted granules including individually listing the subsetted granules, producing zip files for the 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 OffNDAP, the granules needed to be provided in standard formats such as ncDFC or HDF. However, proprietary ASCII or XLS formats were encountered for granules across several datasets. Thus, it was necessary to develop a format translation utility to convert non-standard granules into ncDFC format. This should also be noted that OffNDAP-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.