VIRTUAL COLLECTIONS: An Earth Science Data Curation Service

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

The role of Earth science data owners has traditionally been to maintain central archives that serve as open Earth observing data observatory. However, in order to ensure data are as useful as possible to a diverse user community, Earth science data centers must 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 diffuse 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 unanticipated users. Virtual collections will leverage the amount of time and effort needed to begin research by maximizing certainty of reward and by providing a broadband source of data for unanticipated users. This presentation will define a virtual collection as a product of an Earth science data center 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 follows the general curation model of searching, selecting and synthesizing Earth science data, metadata and information into a single cohesive and useful product (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 criteria and synthesize results.
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

1. Pre-curation framework:
   - Identifying interesting events and the related framing criteria is time consuming. For the SCPEX case study, 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 data in situating the search process.
2. Search:
   - The search for data was limited to data provided by the GHRC to simplify the case study. Limiting data to the GHRC module the search step relatively simple because GHRC provides curation and information from around each field campaign. Searching for data outside the data center requires more time and effort.
3. 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, incomplete spatial information and incomplete cross-collection information. GHRC metadata only provided at the collection level. While the Greenland metadata did include spatial information, the spatial information may not translate the collection level coordinates. Due to these factors, the SCPEX data required that metadata be submitted to the remote selected spatial bounds using `OPendAP`.
4. Synthesize:
   - Data curation: There are many ways to bundle the submitted granules including individual testing the submitted granules, by providing zip files, or 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 `OPendAP`, the granules need to be provided in standard formats such as netCDF or HDF. However, proprietary ASCII or binary formats are encountered for granules across various datasets. Thus, it was necessary to develop a model translation script to convert non-standard granules into netCDF format. This should also be noted in `OPendAP`-based subsetting of the data is only possible if there is a grid type defined for the parameter that was needed for subsetting.
   - System limitations: The Pythoon toolbox could not be directly hosted within the GHRC data center due to security concerns. Therefore, only a static or already executed version of the toolbox is being hosted. This limitation decreases accessibility to the collection and in essence makes the collection non-virtual.

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 also addressed a need for data bundles from the user community. A use case was identified from NASA’s GPM-Diagonal 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 data set 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 interested user community
  - Topic: Snow microphysics
  - Fitness criteria/selection criteria: What is the 3-D structure of falling snow and how does its variability affect remotely sensed retrieval?

Search:
- A search was conducted for databases related to:
  - Microphysical remote sensing or both in-reach and airborne platforms
  - Snow particle size and snow water equivalent measurements also from both ground-based and airborne platforms

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

Conducted by domain experts:
- Metadata were collected from domain experts:
  - 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

Cull results using fitness criteria, spatial and temporal bounds:
- Initial metadata were found for some datasets that were used in the virtual collection.
- Final, relevant publications and reference sources are also listed.

Synthesis:
- Identified data was synthesized with the goal of increasing 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
  - 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 themes/datasets 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.

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/GCPExG5/MULTIPLEDATA101)