VIRTUAL COLLECTIONS:
An Earth Science Data Curation Service

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

The role of Earth science data centers has traditionally been to maintain central archives that serve as open-access Earth observation data. 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 diverse 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 all stakeholders of the data.

Virtual collections will aid in the amount of time and effort needed to begin research by maximizing the return of reward and by providing a more accessibly source of data for unanticipated users. This presentation will define a virtual collection as an Earth science data and will highlight a virtual collection case study created at the Global Hydrology Resource Center data center.

Curation

The steps to curating a virtual collection follows the documentation of a virtual collection. For the SCPE use case, mission reports, campaign blogs, and peer reviewing publications were surveyed to identify an interesting event. Identifying and documenting relevant events is a critical aspect of a field campaign. The search process involved in the virtual collection.

Search:
The search for data is limited to data provided by the GHRC to simplify the use case. Limiting the data to the GHRC made the search step relatively simple because GHRC provides collections around each field campaign. Searching for data outside the dataset 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 only provided at the collection level, and not at the dataset level.
  • Incomplete Spatial Information: Spatial information was only provided at the collection level.
  • Metadata metadata include spatial information, this spatial information matched the collection level coordinates. Due to these requirements, the SCPEs use case required that metadata be provided to the extent collected spatial bounds using OpenDAP.

Synthesis:
• Data Bundling: There are many ways to bundle the submitted granules including individually listing the submitted granules, by providing zip files, or creating bundles. Selecting only one method is totally dependent on the downstream applications that use the bundle.
• Data Format: For dynamic subset of the datasets using OpenDAP, the granules needed to be provided in standard formats such as netCDF, HDF, or ASCII. However, proprietary ASCII or XLS formats were encouraged for granules access between projects. Thus, it was necessary to develop a formal transcription engine 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 subsetting.
• System Limitations: The Python notebook could not be directly hosted within the GHRC data center due to security concerns. Therefore, any 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.

Lessons Learned

Pre-curation framework:
• Identifying interesting events that the related framing criteria is time consuming. For the SCPE use case, mission reports, campaign blogs, and peer reviewing publications were surveyed to identify an interesting event. Identifying and documenting relevant events is a critical aspect of a field campaign.

Search:
• The search for data is limited to data provided by the GHRC to simplify the use case. Limiting the data to the GHRC made the search step relatively simple because GHRC provides collections around each field campaign. Searching for data outside the dataset requires more time and effort.

Select:
• Metadata quality is important and was an limiting factor in selecting data for the virtual collection. Information gaps in the metadata included:
  • Incomplete Temporal Information: Temporal information was only provided at the collection level and not at the dataset level.
  • Incomplete Spatial Information: Spatial information was only provided at the collection level.
  • Metadata metadata include spatial information, this spatial information matched the collection level coordinates. Due to these requirements, the SCPEs use case required that metadata be provided to the extent collected spatial bounds using OpenDAP.

Synthesis:
• Data Bundling: There are many ways to bundle the submitted granules including individually listing the submitted granules, by providing zip files, or creating bundles. Selecting only one method is totally dependent on the downstream applications that use the bundle.
• Data Format: For dynamic subset of the datasets using OpenDAP, the granules needed to be provided in standard formats such as netCDF, HDF, or ASCII. However, proprietary ASCII or XLS formats were encouraged for granules access between projects. Thus, it was necessary to develop a formal transcription engine 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 subsetting.
• System Limitations: The Python notebook could not be directly hosted within the GHRC data center due to security concerns. Therefore, any 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.

Identity Data Center Overview

IDENTIFY DATA CENTER PURPOSE, GOALS, ACCESSIBILITY, USABILITY

DEFINITE FITNESS CRITERIA FOR THE VIRTUAL COLLECTION

SELECT RELEVANT DATA AND INFORMATION

CONDUCTED BY DOMAIN EXPERTS

SEARCH FOR DATA AND INFORMATION

SYNTHESIS

ACCESSIBILITY

USABILITY

Metadata Annotation

Metadata Container

Data Bundle

Contextual Documentation

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 DAACs data holdings and that also addressed a need for data bundles from the user community. A use case was identified from NASA’s GPM (Global Precipitation Measurement) mission’s field campaign data that is archived at the GHRC. The GCPME 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: Benchmarking. What is the 3-D structure of falling snow and how does its variability affect remotely sensed retrieval?

Search:
• A research was conducted for databases related to:
  • Microwave remote sensing on both conical and airborne platforms.
  • Snow particle sizes and snow-water equivalent measurements also from both ground-based and airborne platforms.

Select:
• Data was culled by:
  • Confirming data collection was during the identified temporal period of February 24, 2012
  • Selecting data with parameters most relevant to the fitness criteria:science question
  • Some data was precluded from being selected due to data formatting challenges

Synthesis:
• Identified data was synthesized with the goal of increasing data usability. The GCPME 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 scientific concepts. Micro articles create a knowledge base for users by curating knowledge around the science themes across 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 micro article for the GCPME virtual collection describes the February 24, 2012 event and the science phenomena. This micro article also provides information on the member databases that were subsisted within the virtual collection.
• Finally, the GCPME 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.

https://ntrs.nasa.gov/search.jsp?R=20160014836

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