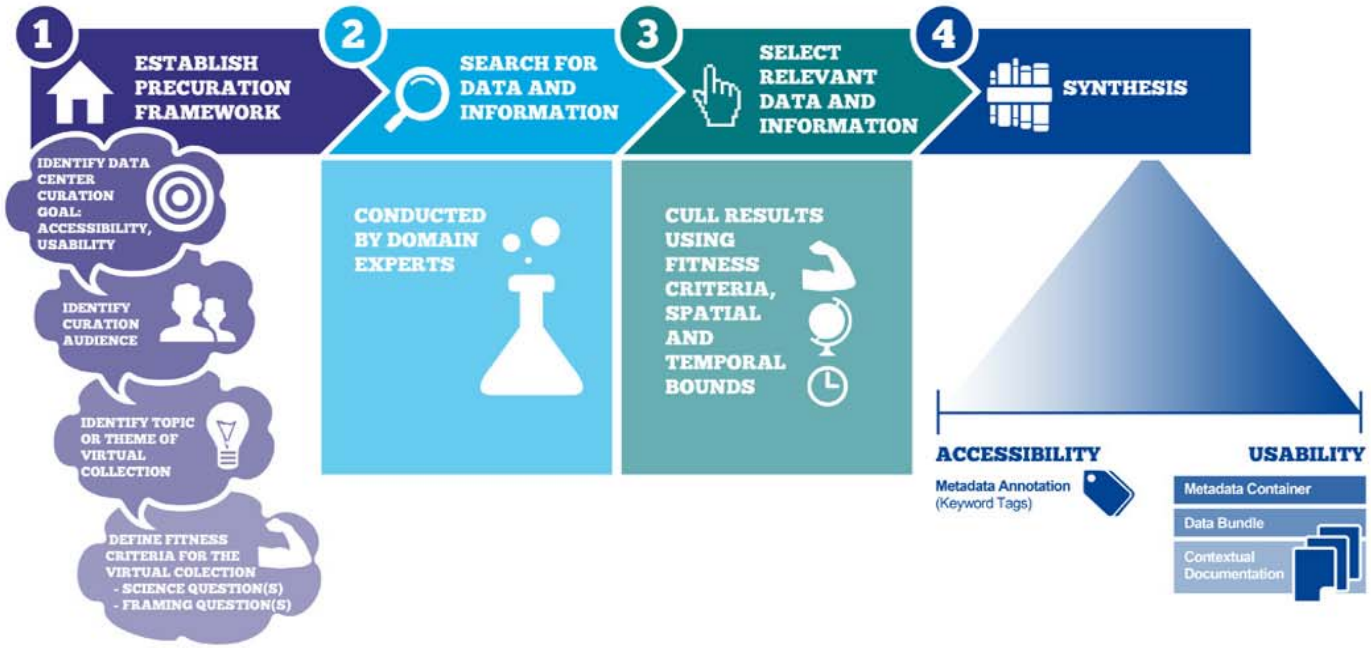
VIRTUAL COLLECTIONS:

An Earth Science Data Curation Service

Kaylin Bugbee[1] (kbugbee@itsc.uah.edu), Rahul Ramachandran[2], Manil Maskey[2], Patrick Gatlin[2]
[1]The University of Alabama in Huntsville, [2]NASA MSFC

The role of Earth science data centers has traditionally been to maintain central archives that serve openly available 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 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 minimize the amount of time and effort needed to begin research by maximizing certainty of reward and by providing a trustworthy source of data for unanticipated users. This presentation will define a virtual collection in the context 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 curating a virtual collection curation follows the geocuration 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, the topic 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 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.

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 Cold-Season 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.

1 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 retrievals?

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

3 Select:
- Data was 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 precluded from being selected due to data formatting challenges

4 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 – iPython notebook for subsetting and selection by parameter (https://github.com/ghrodaac/data-recipes/blob/master/GCPEx-Golden-Case.ipynb)
  - 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 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 GCPEx virtual collection describes the February 24, 2012 event and the science phenomenon driving the event. 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. (https://ghrc.nsstc.nasa.gov/home/micro-articles/snow-microphysics-event-during-gcpex-field-campaign#)

- 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://dx.doi.org/10.5067/GCPEXCS/MULTIPLE/DATA101)**

LESSONS LEARNED

1 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. Identifying and documenting relevant events as a field campaign takes place aids in streamlining the 'search' process.

2 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.

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: Temporal information was provided only at the collection level and in most cases was non-existent at the granule level. This limitation meant that granules had to be manually extracted.
  - Incomplete spatial information: Accurate 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 GCPEx use case required that granules be subsetted to the expert selected spatial bounds using OPeNDAP.

4 Synthesis:
- Data bundling: There are many ways to bundle the subsetted granules including individually listing the subsetted granules, by providing zip files or tar bundles, etc. Selecting one method is totally dependent on the downstream applications that use the bundle.
- Data formats: 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 code 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 iPython 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.

NASA

GHRC
Global Hydrology Resource Center