Geocuration Lessons Learned from the Climate Data Initiative Project

Rahul Ramachandran[1], Kaylin Bugbee[2], Curt Tilmes[3] and Ana Pinheiro Privette[3]

1 NASA/MSFC

[2] University of Alabama in Huntsville

[3] NASA/GSFC

Abstract

Curation is traditionally defined as the process of collecting and organizing information around a common subject matter or a topic of interest and typically occurs in museums, art galleries, and libraries. The task of organizing data around specific topics or themes is a vibrant and growing effort in the biological sciences but to date this effort has not been actively pursued in the Earth sciences. This presentation will introduce the concept of geocuration, which we define it as the act of searching, selecting, and synthesizing Earth science data/metadata and information from across disciplines and repositories into a single, cohesive, and useful compendium.
We also present the Climate Data Initiative (CDI) project as an prototypical example. The CDI project is a systematic effort to manually curate and share openly available climate data from various federal agencies. CDI is a broad multi-agency effort of the U.S. government and seeks to leverage the extensive existing federal climate-relevant data to stimulate innovation and private-sector entrepreneurship to support national climate-change preparedness. The geocuration process used in CDI project, key lessons learned, and suggestions to improve similar geocuration efforts in the future will be part of this presentation.

What is geocuration?

Geocuration is the act of searching, selecting, and synthesizing Earth science data/metadata and information from across disciplines and repositories into a single, cohesive, and useful collection.

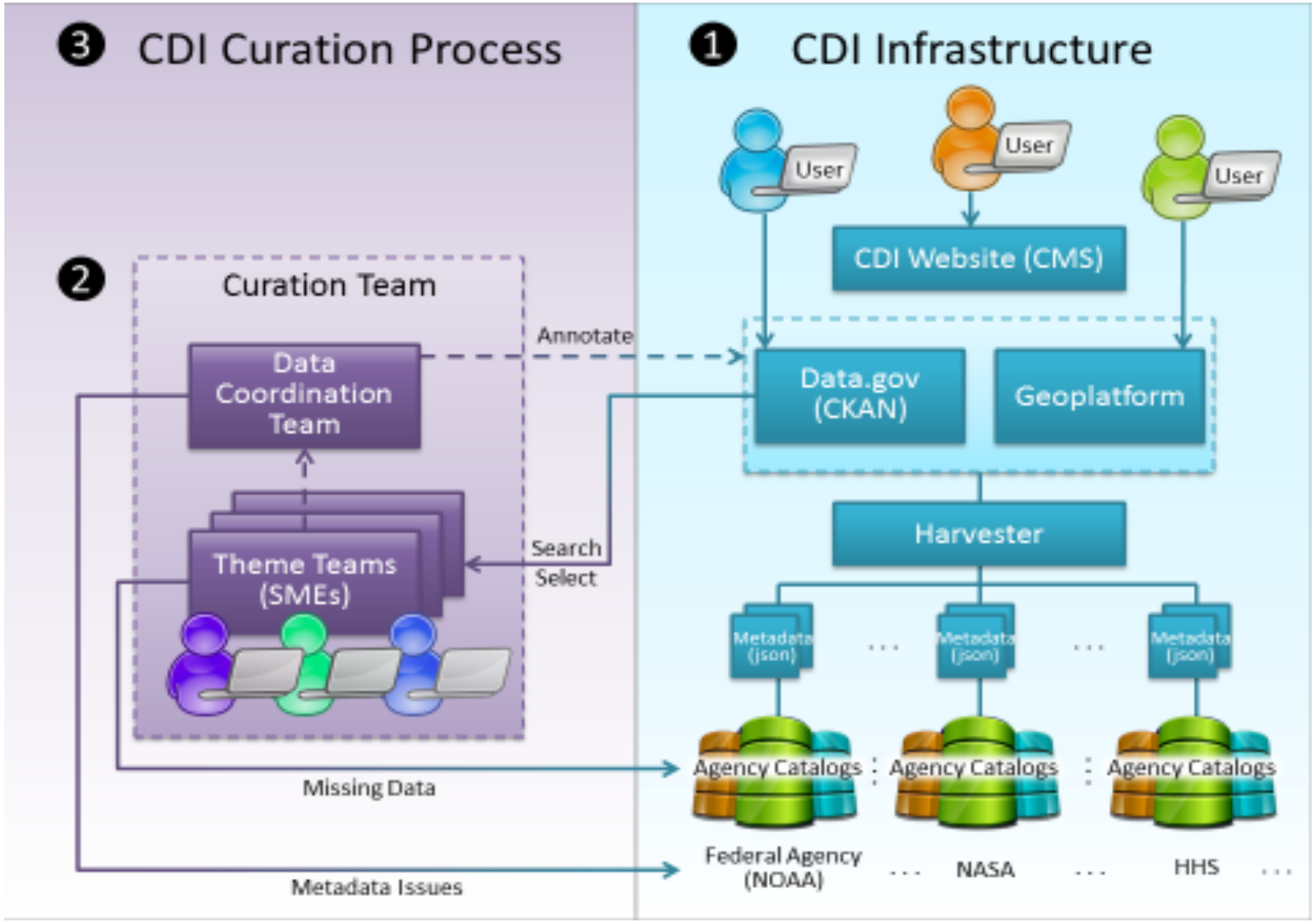
- *Search* step is guided by the cumulative domain expertise of the curators. The collective knowledge of the domain experts is utilized to identify all known relevant data and information resources.
- *Selection* step entails culling the search results based on some "fitness or relevancy" criteria. The fitness criteria can range from simple spatial temporal bounds and resolution, a set of framing questions that define the contextual narrative around the curation effort or fully described use cases.
- Identified data and other information is *synthesized* into a cohesive collection. The consumers of the collection should be able to use the information in his or hers own research or applications with minimal effort.

What is CDI?

- The Climate Data Initiative (CDI) focuses on preparing the United States for the impacts of climate change by leveraging “extensive federal climate-relevant data to stimulate innovation and private-sector entrepreneurship in support of national climate-change preparedness.” (President’s Climate Plan, 16).
- NASA led collaborative effort across federal agencies and scientific disciplines that seeks to make federal climate data both usable and accessible for its defined stakeholders.

CDI Process

1. The data system infrastructure supporting the project
2. The curation team consisting of Subject Matter Experts (SMEs) and informatics experts
3. The curation process itself



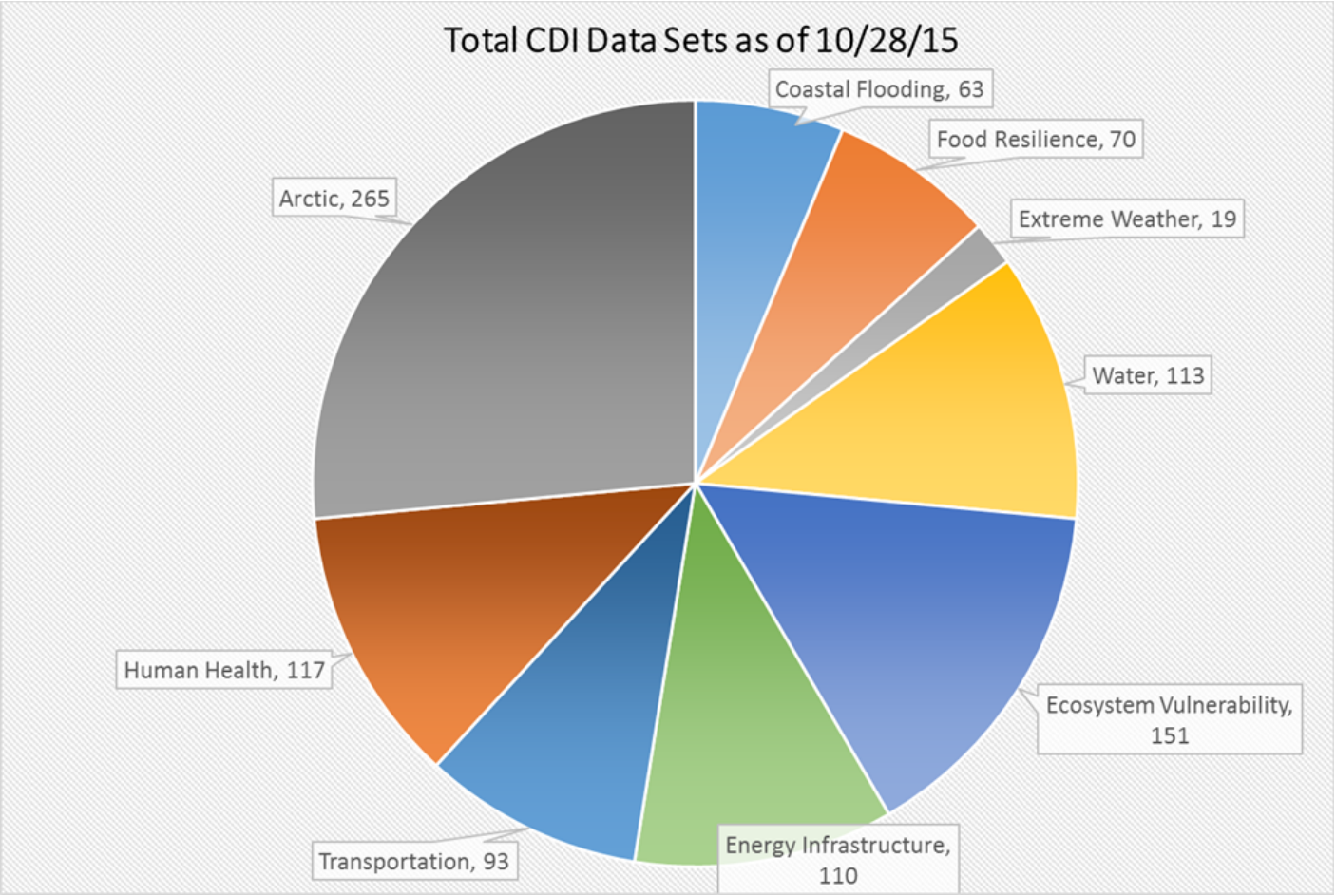
CDI Themes

- Each theme in CDI is incrementally released and is announced by the White House via press releases. The incremental release process for each theme ensures that they are highlighted individually.

Theme	Date Released	Lead Agency
Coastal Flooding	March 2014	NOAA
Food Resilience	July 2014	USDA
Water	November 2014	DOI
Ecosystem Vulnerability	December 2014	DOI
Energy Infrastructure	June 2015	DOE
Transportation	June 2015	DOT
Human Health	April 2015	HHS/CDC
Arctic	September 2015	DOI

Curation Results

- 738 unique datasets
- Over 850 datasets from pre-release theme team submissions were checked for quality by the data coordination team
- Approximately 100 did not pass the metadata quality checks at the time of release.
- CDI website was instrumented with Google Analytics in January 2015 after four of the themes had been released. The numbers from January 2015 are significant. There were around 47,000 unique page views on the CDI site. About 2% of the total visitors browsed the curated data



Challenges faced during CDI Curation Process

1. *Need for Discoverable, Open, and Accessible Data*
   - More often than not, at least one desired dataset by the SMEs on the theme team was not always readily available
2. *Importance of Synthesis*
   - The curated list of data is unable to accurately capture the subject matter experts' intent
3. *Curation is a non-trivial process*
   - Curation for CDI is complicated because of the involvement of many people from multiple agencies using many different infrastructure components and short deadlines for each theme release
4. *Metadata standards help but there are always some issues*
5. *Curation cannot be a one-off activity*
   - Process is dynamic because the curated list changes over time and requires periodic monitoring

Lessons Learned

- Any successful data curation activity (both local and virtual) requires a large pool of open and accessible datasets that are discoverable. Also, metadata catalog(s) play a critical role in enabling successful data curation, especially if the curated data collections are virtual.
- The role of synthesis in curation is often overlooked or glossed over; however, this synthesis often turns out to be an important element in determining the utility of the curated collection. Metadata for the selected data must be synthesized with the intent of curation, captured in a formal structure or information model, and presented to users in a meaningful manner instead of just being presented as a long list of data sets per topic.
- Curation should not be a one-off process. As long as the curated collection is relevant, it requires periodic updates and monitoring to maintain both its quality and value to end users

Lessons Learned (cont)

- There is a need to capture usage metrics because assessing the impact made by the curation effort (Howe et al. 2008) could persuade others of the validity of the process.
- The curation process can be streamlined to encourage continued participation. Transforming the original curators into moderators of the collection instead of just the primary source of content would lighten the burden of curation (Goble et al. 2008).