Computers & Geosciences 88 (2016) 22–29

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Case study

# Climate data initiative: A geocuration effort to support climate resilience

Rahul Ramachandran [a,*], Kaylin Bugbee [b], Curt Tilmes [c], Ana Pinheiro Privette [c]

[a] NASA/MSFC, United States
[b] University of Alabama in Huntsville, United States
[c] NASA/GSFC, United States

ARTICLE INFO

*Article history:*
Received 5 August 2015
Received in revised form
5 November 2015
Accepted 2 December 2015
Available online 5 December 2015

*Keywords:*
Geocuration
Climate data initiative
Climate change
Geoinformatics
Metadata
Virtual collections

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. In this paper, we introduce the concept of geocuration and 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 collection. We present the Climate Data Initiative (CDI) project as a 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. We describe the geocuration process used in the CDI project, lessons learned, and suggestions to improve similar geocuration efforts in the future.

## 1. Introduction

The definition of curation can vary depending on one's perspective. 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. In the library community, the curation process has become more nuanced with the advent of digital content. The digital library community defines curation as "actions people take to maintain and add value to digital information over its lifecycle, including the processes used when creating digital content" (Walters et al., 2011). Similarly, Philip et al. (2004) define curation as the "activity of managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available". A cornerstone component of this curation activity is archiving, whereby selected digital resources are stored and made accessible for future use.

Like the library community, the Earth science data communities also perform curation activities but under the broader umbrella of data stewardship (Peng et al., 2015). These data stewardship activities support the data life cycle by enabling data preservation, accessibility, usability, and sustainability, thereby ensuring quality and reproducibility. The task of organizing data around specific topics or themes is a vibrant and growing effort in the biological sciences (Howe and Yon, 2008) but to date this effort has not earned widespread adoption in the Earth sciences. One reason for this activity gap is that most Earth science repositories have mission statements centered on broad science objectives to support a defined set of science stakeholders around field campaigns, observation platforms and missions. The types of data ingested, archived, published, and distributed must adhere to these guidelines. NASA's Earth Science Data Active Archive Centers (DAACs) are a good example of distributed science repositories (Kobler and Berbert, 1991) with each DAAC's data holdings focused on specific science themes. The data within each repository is aggregated around science projects/missions, instruments or science keywords, and is presented to the user community using this same organizational structure.

There are rapidly emerging causes that drive the need for a finer-grained curation of data and information within Earth science. First, there has been a rapid increase in the growth of the number of Earth science datasets and publications. For example, there are over 14,600 Earth science related data collections (not individual files) available in the Data.gov catalog (Wright, 2014)

[*] Corresponding author.
*E-mail address:* rahul.ramachandran@nasa.gov (R. Ramachandran).

http://dx.doi.org/10.1016/j.cageo.2015.12.002
0098-3004/Published by Elsevier Ltd.

from various U.S. federal agencies.[1] A recent search on Elsevier's journals related to Earth science produced a result of over 40,000 papers published in 2014 alone. Second, the study of Earth as a system has revealed that a specialized focus on one facet of the system does not necessarily capture the dynamics of an interdependent system. Accordingly, research within Earth science has become exceedingly interdisciplinary. This interdisciplinary nature of research requires discovery of both data and information from distributed, multiple domain data and publication repositories.

In this paper, we introduce the concept of *geocuration* and present the Climate Data Initiative (CDI) project (CDI, 2014) as a 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 which seeks to leverage 'extensive federal climate-relevant data to stimulate innovation and private-sector entrepreneurship in support of national climate-change preparedness.' (Climate Action Plan). CDI utilizes Subject Matter Experts (SMEs) from different federal agencies to manually curate data around key climate resiliency themes. CDI exemplifies the need for geocuration given both the complexity of the topic and the types of relevant data available from different federal agencies for climate change. The subsequent sections describe geocuration, the Climate Data Initiative project, the geocuration process used, lessons learned, and suggestions to improve future geocuration efforts.

## 2. 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*. Geocuration is analogous to the concept of verticalization in tool development, where verticalization refers to the customization of a tool (Kohavi et al., 2002) based on a specific science use or domain application. Geocuration serves the same purpose by searching, selecting, and synthesizing data and information based on specific science needs.

Geocuration requires following several systematic steps, each of which serves a specific purpose. The *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. Information resources could include citations for relevant literature, specific workflows, tools, web sites, reports, and documents. The *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. Performing a literature review and identifying relevant data in published journal articles (Karasti et al., 2006) is another approach for selection. Finally, targeting the needs of the intended consumers of the curated collection is another effective way to filter identified information and to determine what needs to be provided by the curation effort (Goble et al., 2008).

Once the selection step is complete, the identified data and other information is *synthesized* into a cohesive collection. The goal of synthesis is to address a set of questions: What has been gathered? Are all the data and information pieces easily identifiable and their associations understandable? Why are these data and information pieces important to the topic? The synthesis should provide a contextual framework for all the gathered information objects. How can this information unit be used? The consumers of the collection should be able to use the information in his or hers own research or applications with minimal effort. The synthesized information can contain data which is stored either locally or virtually and at different levels of granularity. Local data can be aggregated as data bundles containing individual data granules or files. Locally stored data can also be aggregated as a single new product or a file containing curated data parameters from different data sets. On the other hand, virtually stored data can be contained within a virtual collection. A virtual collection is a synthesized collection created from metadata and only includes links to the data's home distributed data repository for final access and use. Virtual collections can have different levels of granularity and can contain individual data files, collection level metadata records or specific data parameters. The ability to create virtual collections using the existing rich metadata catalogs in the Earth sciences offers a promising potential for enhancing data access and use.

There are two approaches to geocuration: manual curation and automated curation. Manual Curation requires Subject Matter Experts to serve as digital librarians, or *geocurators,* who discover and synthesize data and information virtually. One of the main advantages of manual curation is accuracy and trustworthiness to address "suitability of purpose". This is a key requirement for downstream consumers of this curated information, especially in the Earth Sciences. Peng et al. (2015) describes this need by asserting that "…users are asking for data to be dependable in terms of quality and production sustainability, to be from credible, secure, and authoritative sources, to be easily and publicly accessible online." Manual Curation, however, is labor intensive and "a nontrivial undertaking that needs to balance content coverage against content quality" (Goble et al., 2008). Moreover, to be effective, curation needs to become a community activity promoting "collaboration where sheer scale of effort needed can deliver both breadth and economies of scale not possible for each singular participant" (Macdonald and Martinez-Uribe, 2010). Community-driven curation can also provide the editorial oversight to minimize any biases that may occur based on an individual curator's preferences. One example of successful manual curation is described by Howe and Yon (2008) as "biocuration," a topic within the biomedical field, focusing on the activity of organizing, representing and making biological information accessible and usable for specific specialized sub-themes. Biocuration facilitates community-based curation to address the existing gaps in knowledge, provides researchers with a means to quickly find and use massive amounts of complex data quickly, offers insights concerning specific areas of interest and makes it possible to process information faster as data and information is synthesized as part of curation. Extracting, tagging with control vocabularies, and representing data from published literature are the core tasks within biocuration.

Curation is still difficult to achieve in a fully automated manner. There are different approaches and tools that support topic or theme-based searches using text mining or ontological based algorithms (Shamsfard et al., 2006; Yue et al., 2009; Liu, 2010). These approaches by themselves are not enough but can be used as tools to filter down resources that are then manually re-ranked and synthesized (Alex et al., 2008). These tools can support searches across domains and provide automated mediation between different vocabularies used in different repositories to represent similar data (Klien et al., 2001). An example of an automated curation prototype is the "Data Albums" described in Ramachandran et al,. (2014). Data Albums are compiled virtual collections of information related to a specific science topic or an event, containing links to relevant data files (granules) from different

[1] This number does not include other useful publicly available datasets distributed by research laboratories, universities and other organizations.

instruments as well as tools and services for visualization and analysis and information about the event contained in news reports, images, or videos to supplement research analysis. Curation is achieved via an ontology-based relevancy ranking algorithm that filters out non-relevant information and data. We envision that in the near future specialized relevancy ranking algorithms will be able to generate virtual collections for defined curation tasks.

## 3. Climate Data Initiative Project Overview

The President's Climate Action Plan and the Executive Order 13653[2], Preparing the United States for the Impacts of Climate Change, call for the Federal Government to "…develop and provide authoritative, easily accessible, usable, and timely data, information, and decision-support tools on climate preparedness and resilience" to support federal, regional, state, local, tribal, private-sector and nonprofit-sector efforts to prepare for the impacts of climate change. In response to this call, NASA and NOAA were asked to lead the Climate Data Initiative (CDI) and development of a Climate Resilience Toolkit (CRT), respectively. The U.S. Climate Resilience Toolkit (toolkit.climate.gov) provides scientific tools, information, and expertise to help people manage their climate-related risks and opportunities, and improve their resilience to extreme events.[3] 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). It also supports the broader Open Data Policy and integrates this effort with other Open Data Initiatives by adding the new Climate.Data.gov which includes an online catalog of datasets and data products. The Climate Data Initiative is a collaborative effort across federal agencies and scientific disciplines that seeks to make federal climate data both usable and accessible for its defined stakeholders. To date, the CDI and CRT include eight themes, or topics, relevant to climate change resiliency. These themes include Coastal Flooding, Food Resilience, Water, Ecosystem Vulnerability, Human Health, Energy Infrastructure, Transportation and Arctic. Each theme is a curated virtual collection that contains data that is relevant to addressing the challenges of climate resiliency as it relates to a specific aspect of the Earth system and the resulting societal impacts.

Since knowing for whom curation is intended can serve as guide for what curation to provide (Goble et al., 2008), the Climate Data Initiative defined its stakeholders to include decision makers and data innovators. Decision makers are individuals responsible for shaping policy, legislation, finances, social programs, funding, and disaster planning at the national, state, and local levels. These decision makers include policy makers and planners who need to analyze data related to activities that are essential to planning for climate change resiliency. A key data need for decision makers such as GIS analysts, emergency management responders, and natural resource managers is accessible, ready to use data in formats or standard APIs that are supported by a decision support system. Example formats range from KMLs and ESRI's shapefiles to geoTIFFs which can be easily used in geographic information systems.

The CDI is focused on stimulating innovation and entrepreneurship among data innovators in the private sector and the general public who will use data to create and build information and applications for end users. Data innovators are public and private sector software developers that wish to develop new applications that leverage the federal government's openly available climate data. Recognizing that some of the best ideas for government come from outside the government, CDI targets innovators to stimulate the growth of innovative websites, innovative new apps, and other creative tools around the various climate resiliency themes.

## 4. CDI curation process

The three components of the CDI project are: the data system infrastructure supporting the project (Fig. 1, Number 1), the curation team consisting of Subject Matter Experts (SMEs) and informatics experts (Fig. 1, Number 2), and the curation process itself (Fig. 1, Number 3). Fig. 1 provides a bird's-eye view of the CDI curation process and its components.

### 4.1. Curation infrastructure

To curate a virtual collection that includes information about metadata from various agencies across the Federal government, a catalog is required to hold all the metadata in a single repository or location. All federal agencies are mandated to publish metadata for their datasets in the Data.gov (US EOP-OMB, 2009) catalog. Therefore, the Data.gov catalog was the natural choice to serve as the core infrastructure component for the CDI interagency curation effort.

The underlying Data.gov catalog uses the Comprehensive Knowledge Archive Network (CKAN) (Wainwright, 2012) data management system. CKAN is a widely used data management system which makes data discoverable and accessible. It provides tools to streamline publishing, sharing, finding, and using data. Data publishers use CKAN to create a catalog that both describes and makes the data discoverable. Data.gov supports CKAN's open source nature by adding new functionality and customizations as well as repairing CKAN-related bugs. CKAN also provides a RESTful API to programmatically query its catalog, generate statistics, and list datasets by theme.

There are two main types of metadata in Data.gov: geospatial and non-geospatial. All non-geospatial metadata must comply with the Project Open Data (POD) metadata schema. The POD metadata schema is based on Data Catalog Vocabulary (DCAT) and requires JavaScript Object Notation (JSON) format encoding for its records. All agencies provide metadata in POD-compliant JSON files. These metadata records are harvested daily. Validation for schema conformance is performed during the harvest process before the metadata is ingested and published in the Data.gov catalog.

For describing geospatial datasets, the Data.gov catalog supports two types of geospatial metadata standards: ISO-19115:2003 and the Federal Geographic Data Committee's Content Standard for Digital Geospatial Metadata (FGDCCSDGM). Geospatial metadata is typically provided in a Catalog Service for the Web (CSW) endpoint. A mapping, implemented by a crosswalk developed by geospatial metadata experts, is required to transform geospatial metadata to the native Data.gov Project Open Data schema. The crosswalk maps the ISO 19115:2003 metadata into the POD schema. The CSDGM/FGDC metadata is first mapped into the ISO 19115:2003 schema and then subsequently transformed into the POD schema using this same crosswalk.

### 4.2. Curation team

For CDI, geocuration is a manual activity completed by two teams – the theme team and the data coordination team. The

[2] http://www.gpo.gov/fdsys/pkg/FR-2013-11-06/pdf/2013-26785.pdf

[3] https://toolkit.climate.gov/sites/default/files/CRT_factsheet.pdf

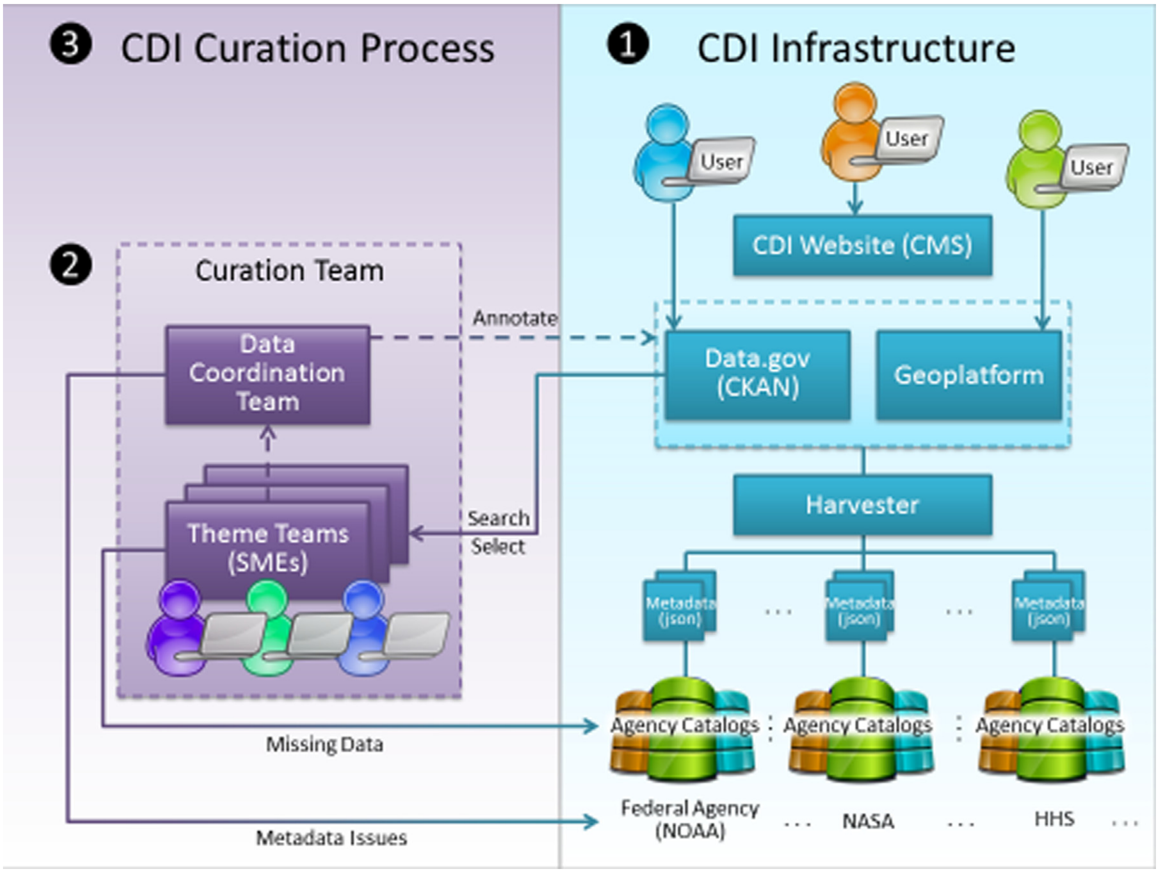


**Fig. 1.** This figure provides an overview of the CDI Curation process andparticipants based on roles and infrastructure components used to publish the final results.

theme team consists of subject matter experts from multiple agencies. The theme team is responsible for recommending sources of authoritative data relevant for a particular climate resilience topic. Each theme team is assigned a team lead and a technical point of contact. The role of the technical point of contact is to liaison with and assist the data coordination team in interacting with different federal agencies in the course of adding missing data to the Data.gov catalog or correcting any metadata issues identified. The data coordination team consists of Earth Science informatics experts with the primary responsibility to check catalog metadata quality, identify problem metadata records, suggest corrections to different agencies to improve metadata quality, and track metrics on data accessibility and usability.

### 4.3. Curation process

Data must meet three criteria to be added to a CDI collection or, as it is known within the CDI, a specific climate resiliency theme. First, a curated dataset should be scientifically relevant to the given climate resiliency topic. The subject matter experts on the theme team ensure that the selected datasets meet this scientific criterion. Second, the curated dataset must be from a reputable source, preferably from a federal agency (in this phase the focus is on federal data resources or data produced under sponsorship of a federal agency). Third, a curated dataset must be accessible and usable. Data accessibility and data usability definitions used by the CDI project are described in detail in the subsequent paragraphs. The data coordination team, with assistance from the theme team and the original metadata providers, is primarily responsible for ensuring that datasets meet these criteria.

The curation process begins with the theme team creating a series of framing questions to guide the selection of datasets that are suitable and relevant to the climate resiliency topic. The theme team uses the Data.gov catalog as a starting point for searching the relevant data for curation. The theme team identifies any missing metadata and notifies the agency producing the data to publish the requisite metadata. The agency producing the data is responsible for providing the metadata to publish and make discoverable in the Data.gov catalog. After the completion of this portion of the curation phase, the theme team gives the data coordination team a list of data and other ancillary information upon which to perform quality checks.

The data coordination team performs quality control checks on the metadata to verify that data is accessible and the associated metadata is robust enough to ensure users can utilize these datasets in their applications. The CDI project defines accessible data as data that is available in convenient and well-known mechanisms that can be easily consumed such as machine application programming interfaces (APIs) or downloadable files in standard formats. Accessible data are sub-divided into data that are directly usable by decision makers and those more suitable for input to tools and applications that an innovator might develop. Accessible data usable by decision makers include data formats that can be readily interoperable with decision support systems such as geographic information systems including ESRI's ArcGIS, and Google Earth. Accessible data usable by innovators includes common data formats that are machine-readable. Machine-readable data are reasonably structured to allow users to write code for automated processing. Machine-readable data provide the most value to innovators by allowing them to quickly reprocess the data or obtain the data automatically in order to populate applications. These types of data can also have APIs to allow innovators to build new tools using these datasets or by bringing together information from various disparate sources.

The quality assessment for all of the metadata in a curated collection is compiled by the data coordination team in a document. This document provides feedback for each individual metadata record and includes all identified issues along with suggestions for improving the records. The responsible data providers within specific agencies are given the feedback document. These quality improvements are performed in an iterative manner. If by chance the metadata corrections are not completed by an agency at the time of the theme release, those datasets are not included in the published theme collection. Since the curation for each theme is an ongoing continuous process, improvements to the metadata records are made after the theme release and new metadata records can be subsequently added to the collection.

## 5. Curation results

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. Additionally, the incremental process encourages users to return to the climate collection, thus creating repeat users. Once a theme is made public, the theme teams are encouraged to continue to add additional datasets to the collection. This ensures the climate themes remain fresh and relevant to returning users.

The user accesses the collection through the main climate page on Data.gov at Data.gov/climate (Fig. 2) or via the Data.gov CKAN API. The pages can be sorted by theme which results in the data collections also being listed by theme. The user can select the 'data' tab to obtain the relevant data catalog listing (Fig. 2). The catalog listing is then displayed in the order of the most recent views where 'recent views' quantifies as the number of views within the last two weeks. Once the user selects a record, information about the dataset is displayed including the agency that provides the data, the spatial extent of the data (if applicable), a short summary about the dataset, and links to access the data (Fig. 2).

To date, eight themes have been released as a part of the Climate Data Initiative (Table 1). These themes were curated by subject matter experts from several Federal agencies, including NOAA, USDA, USGS, and HHS/CDC.

The Climate Data Initiative collection currently consists of 738 unique datasets (Fig. 3). Due to some datasets being included in multiple themes, the number of datasets by theme appears to be higher than the total collection.

The 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.

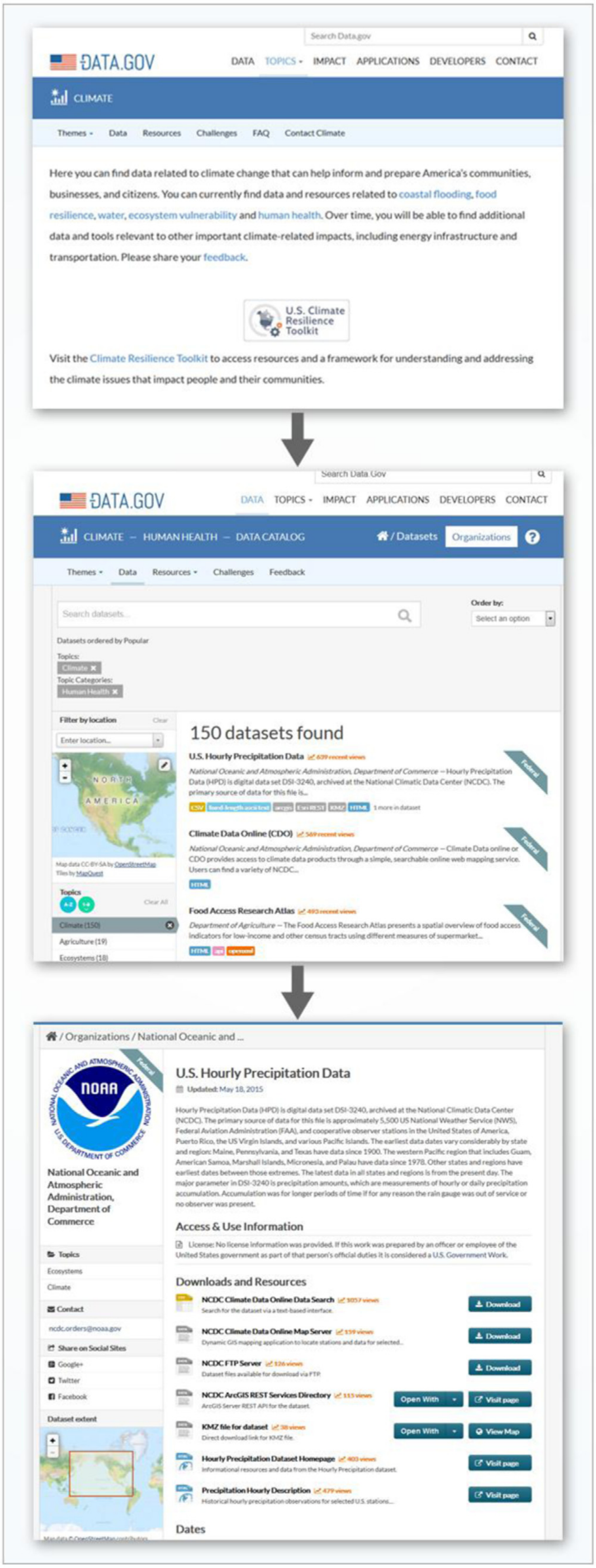
Over 850 datasets from pre-release theme team submissions were checked for quality by the data coordination team. Of these, 738 were made available at the theme release and approximately 100 did not pass the metadata quality checks at the time of release.

## 6. Challenges

Some of the main challenges faced during the CDI curation process are described here:

### 6.1. Need for discoverable, open, and accessible data

Federal agencies are mandated to make their data accessible and to publish metadata in Data.gov. However, more often than



Here you can find data related to climate change that can help inform and prepare America's communities, businesses, and citizens. You can currently find data and resources related to coastal flooding, food resilience, water, ecosystem vulnerability and human health. Over time, you will be able to find additional data and tools relevant to other important climate-related impacts, including energy infrastructure and transportation. Please share your feedback.

Visit the Climate Resilience Toolkit to access resources and a framework for understanding and addressing the climate issues that impact people and their communities.

Datasets ordered by Popular

150 datasets found

**U.S. Hourly Precipitation Data** 639 recent views
*National Oceanic and Atmospheric Administration, Department of Commerce* — Hourly Precipitation Data (HPD) is digital data set DSI-3240, archived at the National Climatic Data Center (NCDC). The primary source of data for this file is...

**Climate Data Online (CDO)** 569 recent views
*National Oceanic and Atmospheric Administration, Department of Commerce* — Climate Data online or CDO provides access to climate data products through a simple, searchable online web mapping service. Users can find a variety of NCDC...

**Food Access Research Atlas** 493 recent views
*Department of Agriculture* — The Food Access Research Atlas presents a spatial overview of food access indicators for low-income and other census tracts using different measures of supermarket...

National Oceanic and Atmospheric Administration, Department of Commerce

U.S. Hourly Precipitation Data

Updated: May 18, 2015

Hourly Precipitation Data (HPD) is digital data set DSI-3240, archived at the National Climatic Data Center (NCDC). The primary source of data for this file is approximately 5,500 US National Weather Service (NWS), Federal Aviation Administration (FAA), and cooperative observer stations in the United States of America, Puerto Rico, the US Virgin Islands, and various Pacific Islands. The earliest data dates vary considerably by state and region: Maine, Pennsylvania, and Texas have data since 1900. The western Pacific region that includes Guam, American Samoa, Marshall Islands, Micronesia, and Palau have data since 1978. Other states and regions have earliest dates between those extremes. The latest data in all states and regions is from the present day. The major parameter in DSI-3240 is precipitation amounts, which are measurements of hourly or daily precipitation accumulation. Accumulation was for longer periods of time if for any reason the rain gauge was out of service or no observer was present.

Access & Use Information

License: No license information was provided. If this work was prepared by an officer or employee of the United States government as part of that person's official duties it is considered a U.S. Government Work.

Downloads and Resources

**NCDC Climate Data Online Data Search** 1057 views
Search for the dataset via a text-based interface.

**NCDC Climate Data Online Map Server** 159 views
Dynamic GIS mapping application to locate stations and data for selected...

**NCDC FTP Server** 126 views
Dataset files available for download via FTP.

**NCDC ArcGIS REST Services Directory** 115 views
ArcGIS Server REST API for the dataset.

**KMZ file for dataset** 38 views
Direct download link for KMZ file.

**Hourly Precipitation Dataset Homepage** 403 views
Informational resources and data from the Hourly Precipitation dataset.

**Precipitation Hourly Description** 479 views
Historical hourly precipitation observations for selected U.S. stations...

**Fig. 2.** The steps to discover a specific curated metadata record for a given theme are presented in the three snapshots. The top image shows the CDI home page. Once a user selects a theme and the data tab, the curated datasets are presented (middle image). The lower image is an example of a specific data set landing page.

**Table 1**
Different climate resilience themes released by CDI.

| 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 |

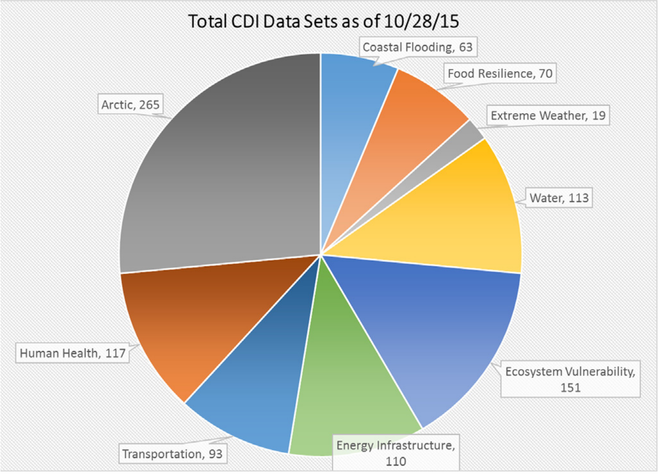


**Fig. 3.** Number of datasets curated under the CDI effort, categorized by the different climate resilience themes.

not, at least one desired dataset by the SMEs on the theme team was not always readily available. The theme teams encountered various challenges when requesting the desired data be added to the Data.gov catalog. These challenges included finding the original data producer, identifying an agency or organization's individual responsible for publishing the metadata into Data.gov, or simply educating the organization on the Data.gov metadata requirements. The theme teams were able to overcome these challenges within their own organizations; however, reaching across agencies sometimes proved difficult.

### 6.2. *Importance of synthesis*

The curated list of data is unable to accurately capture the subject matter experts' intent. While having a curated collection of datasets approved by subject matter experts is valuable, in the end the collection essentially becomes a long directory or a list. Establishing valuable associations between datasets and their intended use is lost in a list. Therefore, the user knows that the datasets in the list have been approved by the subject matter experts but has less certainty when making connections between the various datasets and their possible applications.

### 6.3. *Curation is a non-trivial process*

The process of data 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. Even though a systematic process was designed by the CDI data coordination team and implemented, finding and repairing errors which ranged from missing data sets to broken URLs was an extremely labor intensive effort. This was primarily the role of the data coordination team. As the data coordination team's work progressed, the process of identification and resolution of metadata issues improved. This improvement was due to a better understanding of the Data.gov catalog and their harvesting processes, gained by collaborating with both the Data.gov team and metadata experts from different agencies. This more nuanced understanding of where issues were originating from enabled the data coordination team to provide specific feedback to the theme teams and agencies. Overall, these targeted diagnostics increased the likelihood of metadata records getting improved by the data producers in time for the theme release.

### 6.4. *Metadata standards help but there are always some issues*

Data.gov uses the POD schema to define metadata elements to store in its catalog. However, Data.gov holds metadata for both geospatial and non-geospatial data. Mapping geospatial metadata elements from geospatial standards such as FGDC or ISO 19115 to the POD schema can often be problematic. Two types of errors typically cause the mapping issues. First, lack of obvious one-to-one semantic mappings of certain elements between the two schemas. Second, errors in the software code itself transforming metadata records from one standard to the other.

### 6.5. *Curation cannot be a one-off activity*

Curation cannot be a one-off activity, especially for projects like CDI with ambitious goals and large scope. The curation process is dynamic because the curated list changes over time and requires periodic monitoring. The search and selection process can drive these changes, allowing the curators to discover new relevant data sets that are then added to the relevant theme or topic list. The changes can also be driven by other factors such as data sets no longer being published by the data producer, changes in the infrastructure causing metadata harvesting issues, metadata errors during updates, etc.

The Water theme report (Fig. 4) illustrates these arguments. The initial push of curation by the theme team can be seen leading up to mid-October. During this period, the data coordination team is also checking all submitted metadata records for accessibility and usability. The decline in the number of datasets around the beginning of November illustrates the process of removing all datasets that do not pass quality checks in preparation for the theme release. Notice that the number of associated broken links also decline around this time. Finally, the collection shows continued growth over time as the theme team continues to add new relevant datasets to the collection.

## 7. Discussion

Using subject matter experts to curate data for the climate resiliency themes for the Climate Data Initiative was, overall, a successful endeavor. However, steps can be taken to improve the curation process and resolve some of the issues listed in the section above. Some of the lessons learned from this project that can be applied to any similar curation effort in the future are:

- 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

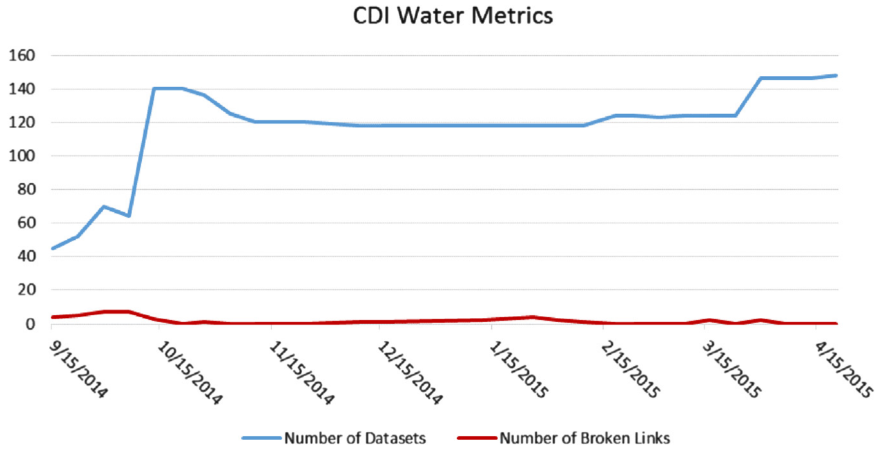


**Fig. 4.** Plot tracking the number of datasets curated under the Water theme over time showing the evolving nature of geocuration.

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. Synthesized information captured in a formal structure such as a graph enables the future users to be able to query and identify different relationships and linkages between different information objects. The Global Change Information System (GCIS) [https://data.globalchange.gov/] is one such example where different information objects such as specific conclusions, figures, data, people etc. contained in the National Climate Assessment report have been synthesized using a rich information model allowing users to trace a specific climate change finding to the attributable papers and identify data sets used to generate specific figures within the report. We envision similar information synthesis in the near future for the CDI project and are currently designing an information model for the Human Health theme. The information model for the Human Health theme will allow users to discover linkages between different datasets and variables that impact health. For example, a user may be able to discover specific climate datasets that impact specific allergens, and find vulnerable populations by locating relevant demographic datasets.

- The use of standards does not eliminate metadata issues, especially if transformations are required between different standards.
- 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.
- 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).
- There is a need to reward or incentivize the curation process. In order to encourage participation, a streamlined citation method for curation efforts would ensure that curators receive recognition for work done. Citation could also potentially encourage the continued use of the curated metadata which could potentially contribute to a longer lifespan for the data.
- 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 methodology followed by the Climate Data Initiative of using both subject matter experts and data experts to curate a collection of climate-related data from across the federal government lends trustworthiness and reliability to the collection. This trustworthiness is essential for decision makers and innovators who wish to plan for climate change resiliency. Additionally, the collaborative nature of the Climate Data Initiative model lays the foundation for future cross-discipline curation efforts in the Earth sciences. The study of Earth as a system has revealed that a specialized focus on one facet of the system does not necessarily capture the dynamics of an interdependent system. The mechanisms of climate change and climate resiliency are similarly interdependent. Better synthesis of the curated data to capture these interdependent relationships is a logical step forward in the pursuit of data discoverability, data accessibility, and ultimately, in the case of the Climate Data Initiative, climate resiliency.

## References

Alex, Beatrice, Claire, Grover, Barry, Haddow, Mijail, Kabadjov, Ewan, Klein, Michael, Matthews, Stuart, Roebuck, Richard, Tobin, Xinglong, Wang, 2008. Assisted curation: does text mining really help? Pacific symposium on biocomputing. Pac. Symp. Biocomput. 567, 556–567.

CDI, 2014. Climate Data Initiative. ⟨http://www.data.gov/climate/⟩.

Goble, Carole, Robert, Stevens, Duncan, Hull, Katy, Wolstencroft, Rodrigo, Lopez, 2008. Data curation + process curation = data integration + science. Briefings Bioinforma. 9 (6), 506–517. http://dx.doi.org/10.1093/bib/bbn034.

Howe, Doug, Yon, Seung, 2008. The future of biocuration. Nature 455 (7209), 47–50. http://dx.doi.org/10.1038/455047a.

Karasti, Helena, Baker, Karen S., Halkola, Eija, 2006. Enriching the notion of data curation in E-Science: data managing and information infrastructuring in the Long Term Ecological Research (LTER) network. Comput. Support. Coop. Work. 15, 321–358. http://dx.doi.org/10.1007/s10606-006-9023-2.

Klien, Eva, Lutz, Michael, Kuhn, Werner, 2001. Ontology-Based Discovery of Geographic Information Services – An Application in Disaster Management Motivating Example : Discovering Services for Estimating Storm Damage in Forests. Compu. Environ. Urban Syst. 30 (1), 102–123.

Kobler B., Berbert J., 1991. NASA Earth Observing System Data Information System (EOSDIS). Digest of Papers Eleventh IEEE Symposium on Mass Storage Systems. doi: 10.1109/MASS.1991.160199.

Kohavi, Ron, Rothleder, Neal J., Simoudis, Evangelos, 2002. Emerging trends in business analytics. Commun. ACM 45 (8), 45–48.

Liu Wei, 2010. Ontology-Based Retrieval of Geographic Information. In: Proceedings of the18th International Conference on Geoinformatics, pp. 1–6. doi: 10.1109/GEOINFORMATICS.2010.5567612.

Macdonald, Stuart, Martinez-Uribe, Luis, 2010. Collaboration to data curation: harnessing institutional expertise. New. Rev. Acad. Librariansh. 16 (S1), 4–16. http://dx.doi.org/10.1080/13614533.2010.505823.

Peng, Ge, Privette, Jeffrey L., Kearns, Edward J., Ritchey, Nancy A., Ansari, Steve, 2015. A unified framework for measuring stewardship practices applied to digital environmental datasets. Data Sci. J. 13, 231–253, February.

Philip, Lord, Macdonald, Alison, Lyon, Liz, Giaretta, David, 2004. From data deluge to data curation. J. Doc. 67 (2), 214–237, doi:10.1.1.111.7425. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=?doi=10.1.1.111.7425.

Ramachandran Rahul, Kulkarni Ajinkya, Maskey Manil, Bakare Rohan, Basyal Sabin, Li Xiang, 2014. Data albums : an event driven search, aggregation and curation tool for earth science, In: Proceedings of the IEEE Geoscience and Remote Sensing Society. Quebec City, Canada.

Shamsfard, Mehrnoush, Nematzadeh, Azadeh, Motiee, Sarah, 2006. ORank : an ontology based system for ranking documents. Int. J. Comput. Sci. 1 (3), 225–231.

U.S. EOP-OMB. 2009. M-10-06: Open Government Directive. ⟨http://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-06.pdf⟩.

Wainwright, Mark. 2012. Using CKAN : Storing Data for Re-Use. In OR2012: Open Repositories. ⟨http://ckan.org/files/2012/08/OKF-OR12-poster.pdf⟩.

Walters, Tyler, Skinner, Katherine, Libraries Association of Research, 2011. New Roles For New Times: Digital Curation For Preservation. Association of Research Libraries.

Wright, Forrest, 2014. Data gov. J. Bus. Financ. Librariansh. 19 (1), 77–82. http://dx.doi.org/10.1080/08963568.2014.855090.

Yue, Peng, Gong, Jianya, Di, Lianlian He, Liping, Wei, Yaxing, 2009. Integrating semantic web technologies and geospatial catalog services for geospatial information discovery and processing in cyberinfrastructure. GeoInformatica 15, 273–303. http://dx.doi.org/10.1007/s10707-009-0096-1.