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.

Published by Elsevier Ltd.

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 non-trivial 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.