Computers and Geosciences 112 (2018) 76–82

Contents lists available at ScienceDirect

Computers and Geosciences

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

The art and science of data curation: Lessons learned from constructing a virtual collection

Kaylin Bugbee [a,*], Rahul Ramachandran [b], Manil Maskey [b], Patrick Gatlin [b]

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

ARTICLE INFO

Keywords:
Virtual collections
Data accessibility
Data usability
Atmospheric science
Geocuration
Earth science

ABSTRACT

A digital, or virtual, collection is a value added service developed by libraries that curates information and resources around a topic, theme or organization. Adoption of the virtual collection concept as an Earth science data service improves the discoverability, accessibility and usability of data both within individual data centers but also across data centers and disciplines. In this paper, we introduce a methodology for systematically and rigorously curating Earth science data and information into a cohesive virtual collection. This methodology builds on the geocuration model of searching, selecting and synthesizing Earth science data, metadata and other information into a single and useful collection. We present our experiences curating a virtual collection for one of NASA's twelve Distributed Active Archive Centers (DAACs), the Global Hydrology Resource Center (GHRC), and describe lessons learned as a result of this curation effort. We also provide recommendations and best practices for data centers and data providers who wish to curate virtual collections for the Earth sciences.

1. Introduction

Museums, art galleries and libraries interpret and provide context for objects through curation. Curators, or subject matter experts, manage the curation process and are responsible for selecting relevant objects to the collection. Within the library sciences, collections typically fall into two categories: general collections and special collections. Special collections are "generally characterized by their artifactual or monetary value, physical format, uniqueness or rarity, and/or an institutional commitment to long-term preservation and access" (Dooley et al., 2010). Museums, galleries and libraries specifically separate special collections from general collections making access to special collections sometimes limited or restricted. However, with the advent of digital libraries and archives, the special collection concept has expanded and become more flexible. Digital archives permit objects to exist in multiple collections and to be organized in various ways simultaneously (Geisler et al., 2002). Digital collections can be created dynamically, can exist for a given amount of time (Geisler et al., 2002) and, unlike physical special collections, can be available to any user. These digital collections, also known as virtual collections, are curated around a topic, an organization or for a specific purpose (Geisler et al., 2002) and provide a service beyond archiving and disseminating information.

Earth science data centers are also in need of providing services beyond simply archiving and distributing data. A virtual collection is an Earth science data stewardship activity that improves "the information content, accessibility, and usability of data and metadata" (Peng et al., 2015). However, curating Earth science data is more nuanced than curating library resources because selecting data requires an understanding of the data's context and how that data should and will be used. Previously, a virtual collection has been defined as "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" (Ramachandran et al., 2016). We define a virtual collection as an end product of a curation activity that searches, selects and synthesizes diffuse data and information resources around a specific theme, topic or event.

For the Earth science data community, there are many motivations for creating virtual collections. Primarily, virtual collections add to the value of data by increasing its use (Uhlir, 2010). Virtual collections encourage greater data use by increasing the analytic potential of data. Palmer et al. define analytic potential as "the likelihood that a data set will be of value for future analysis by others, not just for replication, but also for new applications" (2012). Virtual collections increase this likelihood by highlighting relevant data that provide "'satisfactory answers' to the

* Corresponding author.
E-mail address: kbugbee@itsc.uah.edu (K. Bugbee).

https://doi.org/10.1016/j.cageo.2017.11.021
Received 23 March 2017; Received in revised form 15 November 2017; Accepted 22 November 2017
Available online 24 November 2017

questions raised by actual and potential users" (Hjørland, 1998) and by providing a context for that data. Increased analytic potential is especially significant for data generated in the long tail of science. The long tail of science consists of smaller research projects and accounts for up to 80% of all scientific research (Heidorn, 2008; Palmer et al., 2011) conducted. Long tail science data has the "potential for analysis across aggregates, similar to big science. But their value for reuse may also be complementary, as a unique piece of a complex puzzle or an important addition to a series of measures over time" (Palmer et al., 2011). Virtual collections are uniquely positioned to highlight data relevant to case studies and other events important to long tail data analysis.

In addition to improving data usability, virtual collections also increase analytic potential by improving the discoverability and accessibility of the data. Improved discoverability is achieved by reducing barriers to entry introduced by the limitations of current data systems. For example, NASA's Earth science data is archived and distributed by twelve Distributed Active Archive Centers (DAACs). Each DAAC has a science mission and data is assigned to an archive center based on the data's alignment with the DAAC's science mission. An active assumption is that users are aware of each DAAC's science area yet data discovery is confusing for users that are unaware of the DAAC's science mission. Once data is assigned to a DAAC, that data is organized and archived around both the instrument that collected the data and the related science parameters. Data is also sometimes organized around the field campaign during which the data was collected. For example, Advanced Microwave Scanning Radiometer (AMSR) brightness temperature observations are used to create a wide variety of science data products such as precipitation rate, sea ice concentration and sea surface temperature at various processing levels. The DAACs then organize these science data products by processing level or the generated science parameter. This organizational structure suits the initial user community (Parsons and Duerr, 2005) or the single designated community by providing services that satisfy the "needs of a designated community through an understanding of that community's knowledge base" (Palmer et al., 2011). Once these data become publically available, it is often assumed that new users will resemble the initial user community and have a similar knowledge background. This assumption, however, can alienate or overlook potential unanticipated users who may want to use the data. As Earth Science becomes increasingly interdisciplinary, moving beyond traditional boundaries in order to study Earth as an integrated system, additional unanticipated users will search for and use unfamiliar data. These unanticipated users have a need to understand the context or appropriateness of use of the provided data (Parsons and Duerr, 2005) or else risk misinterpretation on appropriateness of use if the context is not understood. Similarly, unanticipated users need to understand whether data is fit for an identified purpose. Fit for purpose is concerned with the "data quality for the intended use" (Palmer et al., 2011). For example, some data may be appropriate for answering broad science questions but would not be fit for answering questions that require detailed observations. Additionally, there is a need to transfer knowledge claims (Zimmerman, 2007), informal knowledge and background information to both unanticipated users and the initial user community. Documentation, informal communication and metadata attempt to capture this knowledge (Hjørland, 1998), but this information is often crafted by subject matter experts who are familiar with the data and can therefore be difficult to interpret for users outside the designated community. In order to add context, transfer knowledge and ease barriers to discovery and use, trustworthy data and information should be provided for unanticipated users. The curation of virtual collections add context and transfer knowledge by anticipating the science needs of both the initial user community and unanticipated users. Curated virtual collections serve as an authoritative source of data and information for scientific researchers not as familiar with datasets outside their area of expertise. Virtual collections also map the science needs to the data and parameters, which provides an indication as to whether the data is fit for a given purpose. Virtual collections attempt to accommodate "the broadest possible use of including unanticipated use of the data while discouraging data misuse" (Parsons and Duerr, 2005). We envision, in the near future, that data centers will provide value added services to their user communities by creating virtual collections focused around related science themes.

More broadly, virtual collections can ease discovery of and access to data and information that pertain to a common theme but are dispersed and managed separately across multiple data centers. The current distribution system often requires manual searches across disparate data center portals. To streamline the search process within Earth observation data, NASA has created the Earthdata Search portal in order to bring together the metadata records of all twelve DAACs into one access mechanism. We envision that data and information from various sources will be curated in an environment such as the Earthdata Search portal in order to create virtual collections that utilize data from multiple data centers. These cross-curated collections could serve various communities, including unanticipated users, by bringing together data to assist in case study research, disaster analysis and recovery, and cross-disciplinary research problems.

This paper presents our experience in publishing a virtual collection at the Global Hydrology Resource Center (GHRC) as a part of our core data services. We will outline a broad methodology for curating a virtual collection and, leveraging a specific use case, will systematically describe our approach to searching, selecting and synthesizing data and information for a virtual collection. We will conclude by highlighting lessons learned as a result of this experience and will discuss best practices for data providers and archive centers that wish to create virtual collections. As we have learned, creating a virtual collection is a nuanced process that requires a systematic approach formulated around a rigorous scientific question to guide the curation process. Communicating that scientific question and the associated knowledge via both the curation effort and the included contextual documentation distinguishes the virtual collection from a traditional dataset, and therefore makes the virtual collection a necessary value added service for Earth science data centers.

2. Methodology

The steps to curating a virtual collection (Fig. 1) follow 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). The virtual collection curation process is formulated from the search perspective and assumes that a user is either searching for relevant data to test a pre-formulated hypothesis or is searching for a new data problem to explore. Before beginning the search process, the curators define the framework around which the virtual collection will be built (Fig. 1, Step 1). This framework ensures a cohesiveness to the collection and includes defining the goal, audience, topic and fitness criteria of the virtual collection. Curation goals are defined by the data archive and should address the archive's needs of improving data discovery, data accessibility and/or data usability. The curation goal should also take into account the data center's available resources and capabilities. Next, defining the targeted audience for the intended virtual collection and assessing the collection's value to the data center's users is essential since "the quality of collection development is related to the ability to meet the requirements of the users and supply them with satisfactory answers" (Hjørland, 1998). The virtual collection should target one of two core audiences: domain experts that are already familiar with the data or unanticipated users such as interdisciplinary researchers, students, decision makers and data innovators. After the curation goal and audience have been chosen, a topic, theme or event around which the virtual collection will be curated is identified. The topic should relate to the data archive's science mission statement and should be relevant to the data archive's stakeholders. Finally, fitness criteria are defined to guide the selection of data and information for the virtual collection. Since "the acquisition of data is guided by a doable research problem" (Zimmerman, 2007), formulating a science question is one method for defining fitness criteria. The science question can either