ISSN 2161-3974

JeSLIB 2017; 6(2): e1117
doi:10.7191/jeslib.2017.1117

Journal of eScience Librarianship
putting the pieces together: theory and practice

Full-Length Paper

The Evolution, Approval and Implementation of the U.S. Geological Survey Science Data Lifecycle Model

John L. Faundeen and Vivian B. Hutchison

U.S. Geological Survey, Reston, VA, USA

Abstract

This paper details how the U.S. Geological Survey (USGS) Community for Data Integration (CDI) Data Management Working Group developed a Science Data Lifecycle Model, and the role the Model plays in shaping agency-wide policies and data management applications. Starting with an extensive literature review of existing data lifecycle models, representatives from various backgrounds in USGS attended a two-day meeting where the basic elements for the Science Data Lifecycle Model were determined. Refinements and reviews spanned two years, leading to finalization of the model and documentation in a formal agency publication[1].

The Model serves as a critical framework for data management policy, instructional resources, and tools. The Model helps the USGS address both the Office of Science and Technology Policy (OSTP)[2] for increased public access to federally funded research, and the Office of Management and Budget (OMB)[3] 2013 Open Data directives, as the foundation for a series of agency policies related to data management planning, metadata development, data release procedures, and the long-term preservation of data. Additionally, the agency website devoted to data management instruction and best practices (www2.usgs.gov/datamanagement) is designed around the Model's structure and concepts. This paper also illustrates how the Model is being used to develop tools for supporting USGS research and data management processes.

1 The United States Geological Survey Science Data Lifecycle Model http://pubs.usgs.gov/of/2013/1265
2 OSTP https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf
3 OMB https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2013/m-13-13.pdf

Correspondence: John Faundeen: faundeen@usgs.gov
Keywords: data management, lifecycle model, USGS, geology

JeSLIB 2017; 6(2): e1117
doi:10.7191/jeslib.2017.1117

Introduction

The USGS has recognized that a more focused approach to data and information issues is needed to ensure the agency's science products are adequately preserved and easily accessible. In 2009, the USGS established the Community for Data Integration (CDI) to leverage existing expertise throughout the agency and through partnerships in data management, information technology, records and information management, and information governance. Within CDI, the Data Management Working Group (DMWG) seeks to bring cohesion around data management best practices and workflows in the agency. One of the notable outputs of the DMWG is the conceptualization, development, and publication of a data lifecycle model, from which agency data management activities are being framed. The Science Data Lifecycle Model has proven useful to the USGS in a variety of ways. A data management website is designed around the Model, serving agency scientists with best practices and tools for each component of the Model. A suite of data management policies was issued in 2017 that directly address elements of the Model. Finally, the Model is used to build and support enterprise applications and tools related to data management for agency scientists. The Science Data Lifecycle Model is a critical component of communication and practice of data management in the USGS.

Beginnings

Initially, the DMWG was confident that an applicable model could be identified, perhaps minimally refined, and then adopted for USGS use. Literature searches conducted by the cross disciplinary group revealed many lifecycle models of potential relevance. Potential models were acquired for review using a variety of methods of search and retrieval to locate the models from sources including universities, other Federal agencies, and international organizations. Weekly meetings ensued, held for several months, to review each model in depth. For each model, the team discussed each element in the perspective lifecycle and determined its ability to represent the workflows of the USGS from a variety of discipline perspectives. Originally, the team had envisioned finding a model that would address the agency needs; however, while several models had useable elements, no single model captured the data management requirements of USGS scientists completely. Soon, it became apparent that only portions of a given model directly addressed the processes and workflows by which USGS scientists develop and manage their research data. The pertinent elements were noted, as well as elements missing. Below are examples of various models reviewed:

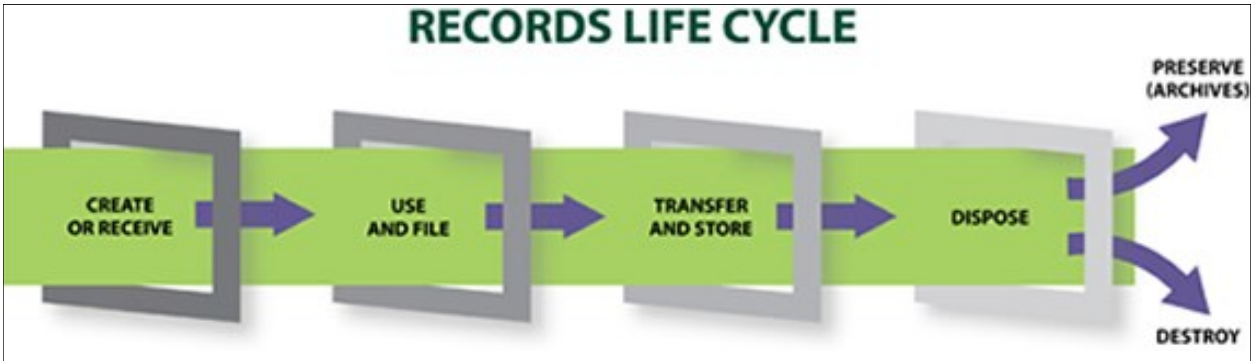

Figure 1: Michigan State University Records Life Cycle (http://archives.msu.edu/records)

JeSLIB 2017; 6(2): e1117
doi:10.7191/jeslib.2017.1117

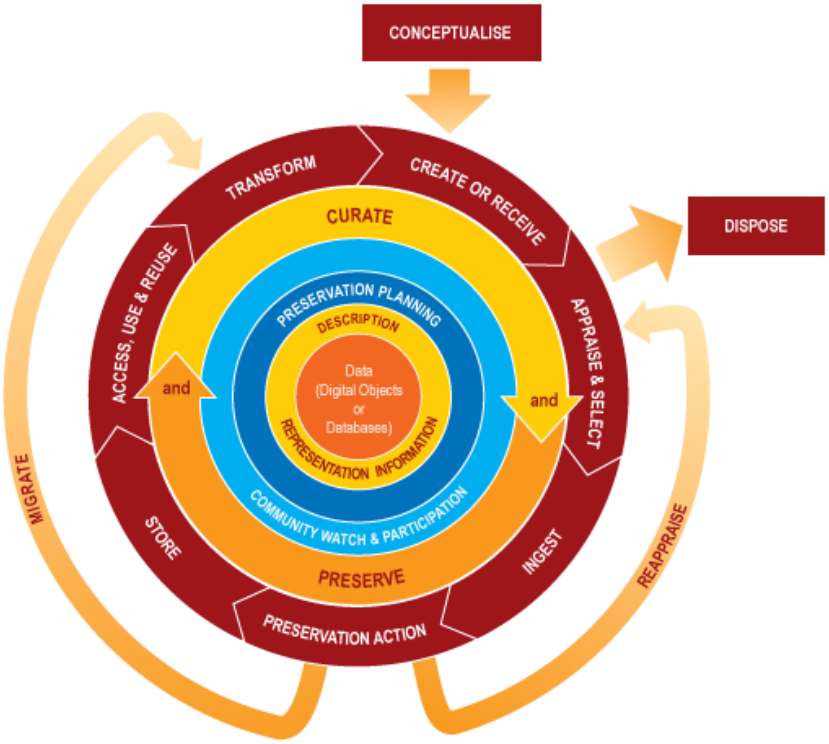

Figure 2: Digital Curation Centre Lifecycle Model (http://www.dcc.ac.uk/resources/curation-lifecycle-model)

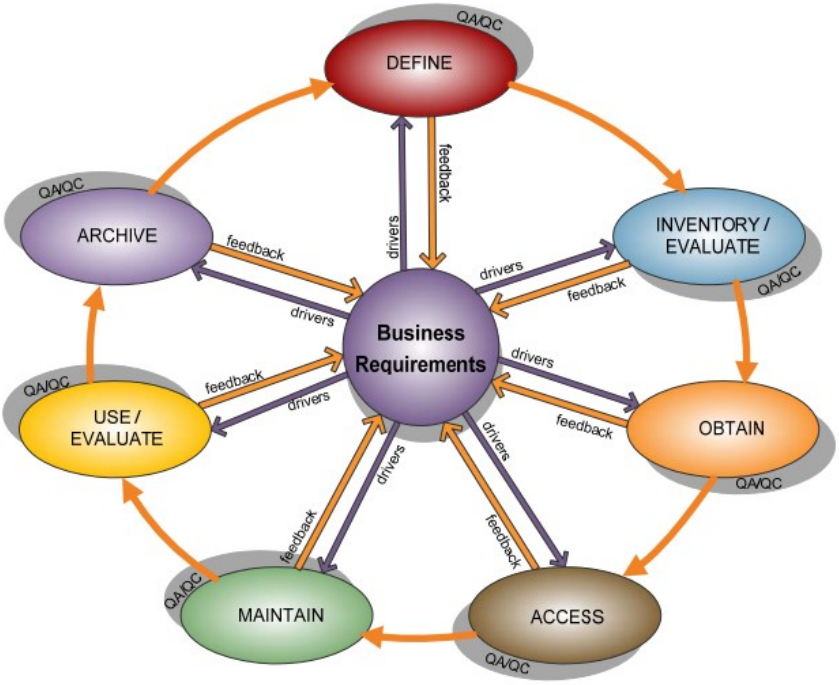

Figure 3: FGDC Stages of the Geospatial Data Lifecycle (https://www.fgdc.gov/policyandplanning/a-16/stages-of-geospatial-data-lifecycle-a16.pdf)

JeSLIB 2017; 6(2): e1117
doi:10.7191/jeslib.2017.1117

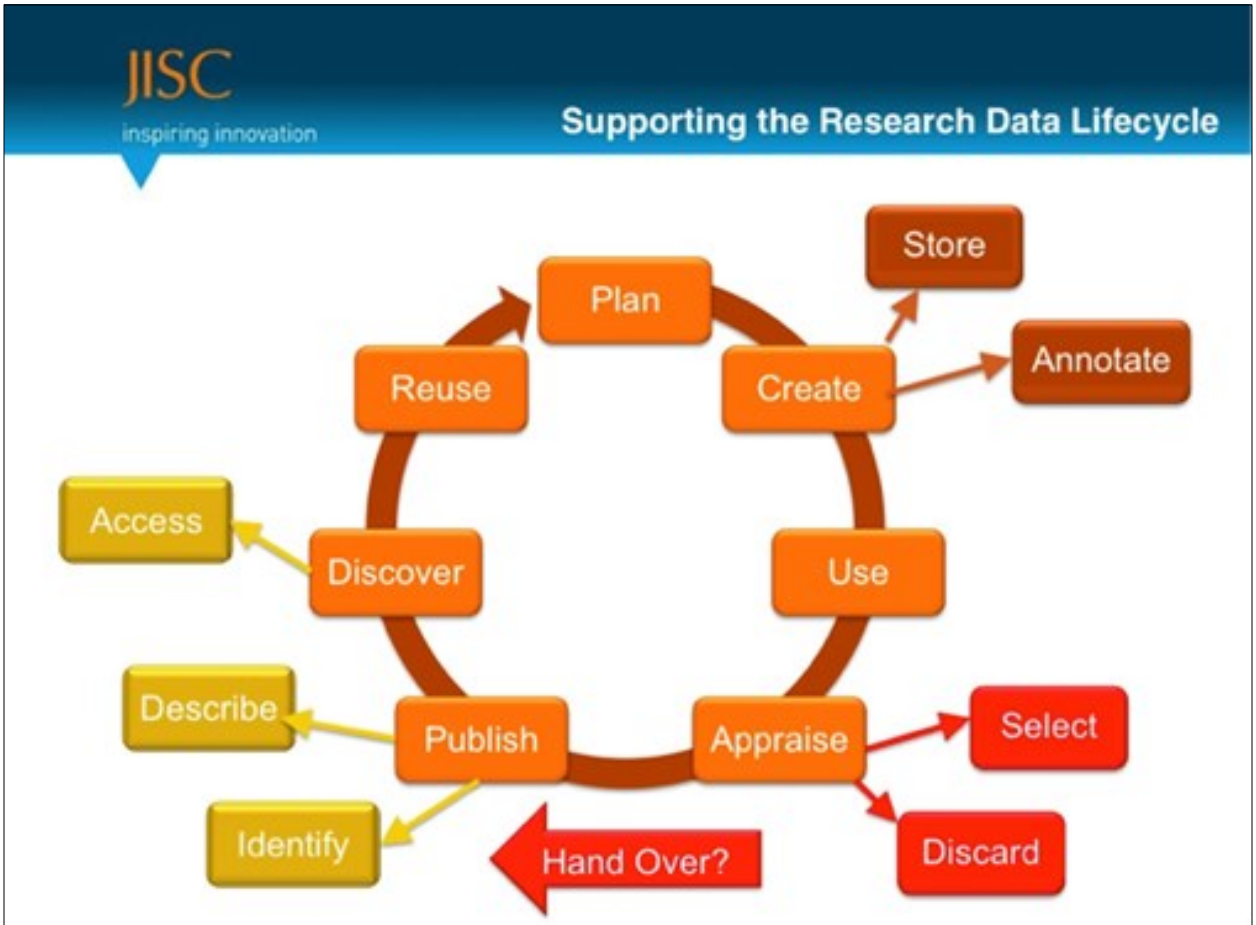

Figure 4: JISC Webinar, Simon Hodson, Programme Manager, Managing Research Data, Digital Infrastructure Team (http://www.jisc.ac.uk/media/documents/events/2011/10/Research%20Data%20Challenge%20presentation.pdf)

Having examined and analyzed over 50 science data lifecycle models, the USGS decided to develop its own interpretation of the lifecycle. There were many reasons for this decision, such as attempting to reduce the complexity observed in many models, as well as focusing on the primary work stages USGS scientists need to address as they compile and execute data management plans. Often, the models reviewed contained steps or elements that did not align well with the scientific workflows or processes in USGS. When developing the resulting Model, the team invested time in determining its graphical flow. In drafts, this varied from linear to circular. The team recognized the importance of illustrating potential circular paths of certain elements, and a linear continuum that ultimately results in the release of a science data product. Determining how intuitive the flow appeared became a major effort that spanned many hours and various iterations before the group could agree on a new model.

The result was “The USGS Science Data Lifecycle Model,” an illustration with accompanying explanation that has, since publication in 2013, broadly influenced the direction of science data management in the agency. It is important to note the placement of the Preserve element in the Model. Preserve is strategically located in the lifecycle before Publish/Share, as a means of

JeSLIB 2017; 6(2): e1117
doi:10.7191/jeslib.2017.1117

emphasizing the importance of taking steps to protect one of the agency's most valuable assets: science data. Also noteworthy is the linear illustration of the Model. While the complexity of the science research process could easily be represented with many circular figures, the authors wanted the Model to be quickly understood and thus opted for the simplicity of a linear graphic.

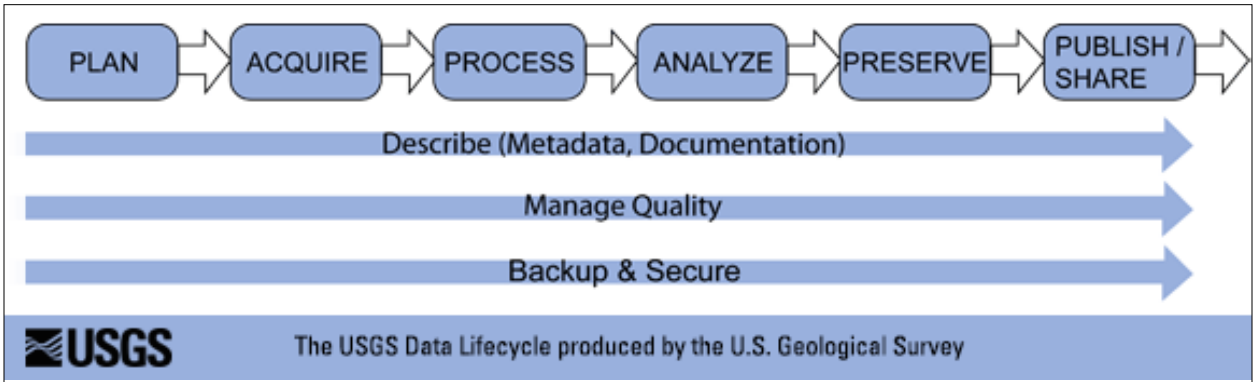

Figure 5: USGS Science Data Lifecycle Model.

Science data are considered one of the most important corporate assets of the USGS. The agency recognizes that the result of well-managed data creates efficiencies in how science is done, improves provenance in the science iteration process, supports scientific review and integrity, allows for reproducibility, and maximizes the effective use and value of data. The USGS is using the Model as a framework for policies and workflows, to ensure scientists are managing data through their lifecycle, from project inception to data release and long-term preservation.

In 2017, the USGS published four new policies[4] around data management activities: Scientific Data Management Foundation; Metadata for USGS Scientific Information Products Including Data; Review, Approval, and Release of Information Products; and Preservation Requirements for Digital Scientific Data. These policies were introduced where aspects of data management were either never addressed or were not sufficiently stated in the USGS Survey Manual, which instructs scientists and others of requirements and responsibilities with science data. Further, the USGS policies were developed to support a series of Open Data initiatives introduced during the Obama Administration. The USGS policies directly reflect various elements of the Science Data Lifecycle Model, thus emphasizing the influence that the Model has on the policy process within the agency.

The Scientific Data Management Foundation policy instructs USGS scientists to address the entirety of the Model. It specifies, additionally, that each new science project must have a data management plan. In the Model, the Plan concept is first. Data management planning encourages the scientist to consider all elements of the Model in the beginning phases of the project.

4 Foundation Policy https://www2.usgs.gov/usgs-manual/500/502-1.html; Metadata for USGS Scientific Information Products Including Data https://www2.usgs.gov/usgs-manual/500/502-7.html; Review, Approval, and Release of Information Products https://www2.usgs.gov/usgs-manual/500/502-4.html; Preservation Requirements for Digital Scientific Data https://www2.usgs.gov/usgs-manual/500/502-9.html

JeSLIB 2017; 6(2): e1117
doi:10.7191/jeslib.2017.1117

The Metadata for USGS Scientific Information Products Including Data policy introduces a requirement for metadata that sufficiently describes data or information products such that they can be understood by other scientists and reused in future science research. Metadata are required to be complete, follow internationally recognized standards, include a digital object identifier, and be submitted to the USGS Science Data Catalog, an online application that increases access to USGS science data. The Metadata element in the Model, on which the policy is based, are included in "Describe" and are portrayed as a continuous line, representing an activity that should be done throughout the project timeline.

The Review, Approval, and Release of Information Products policy specifies the need for a digital object identifier for data, and outlines requirements, such as peer reviews for data and metadata, which further ensure the integrity of the data published by the USGS. This policy reflects several aspects of the Model. The requirement for a digital object identifier (DOI) is associated with both the Preserve and Publish/Share elements in the Model. The DOI is used largely for data citation purposes and to provide a persistent, online address to a data resource, regardless of where it physically resides over time. The element Preserve is directly stated in the language of the policy as part of the review and approval requirements; "Scientific data released by the USGS must be managed and distributed through a data system that can ensure the long-term preservation, discoverability, accessibility, and usability of the resource."

Finally, the Preservation Requirements for Digital Scientific Data policy addresses a requirement for USGS data to be properly preserved, such that they can be accessed in the future. The policy specifically identifies critical aspects to successful data preservation such as multiple copies and storage locations, data viability and integrity, information security, metadata, and file formats. This policy language directly relates to the Model element Preserve, which refers to storing data for long-term access and use. This is an important aspect of data management activities as it allows for future reuse of the science data.

The publication of these policies elevates the critical importance of the Model for USGS. The Model creates a framework upon which USGS policy is based. Beyond the policies, the USGS has additionally utilized the Model to frame useful implementation tools for data management and educational products.

USGS Data Management Website

The USGS Data Management Website[5], launched in 2012 as a result of collaborative work funded through the CDI, was originally conceived around the concepts of the Science Data Lifecycle Model. Designed as a mechanism to provide guidance, tools, and best practices around each concept of the Model, the Website also supports USGS policy by describing ways to implement the policy requirements.

The Website navigation is built directly from the USGS Scientific Data Lifecycle Model, illustrating another critical use of the Model in shaping the way in which data management is incorporated in USGS science workflows and culture. Since the Data Lifecycle process is complementary to the research process, it enables researchers to align data management requirements at the appropriate stages in their individual, respective projects. Users can access each element of the Model to uncover more detailed topics (e.g., create a Data

5 Supporting and Enabling USGS Data Management http://www2.usgs.gov/datamanagement/index.php

JeSLIB 2017; 6(2): e1117
doi:10.7191/jeslib.2017.1117

Management Plan), and easily locate Best Practices, Key Points, Tools, and Recommended Readings. Tying the Website to the Model helps users learn how each element fits in the scheme of data management, and how it may be applied.

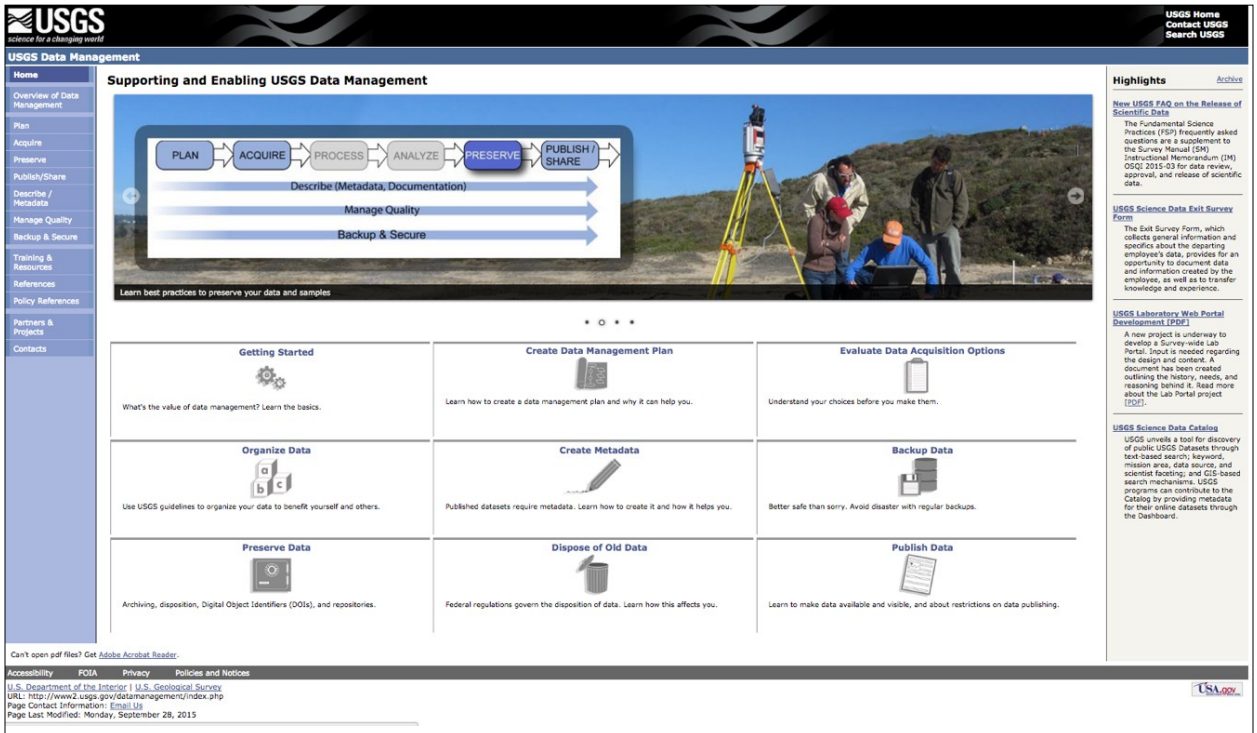

Figure 6: USGS Data Management Website (www2.usgs.gov/datamanagement)

Tools Supporting the Science Data Lifecycle

An additional way in which the U.S. Geological Survey is reinforcing its agency scientists' ability to perform critical science data management activities is through the development or support of online tools and applications that align to the fundamental principles represented in the Science Data Lifecycle Model.

Some examples of data management tools available to scientists and data managers include:

DMPTool: https://dmptool.org
The Data Management Planning Tool allows scientists to collaborate on the creation of a data management plan through a step-by-step process. The DMPTool, developed and supported by a consortium of partners including the California Digital Library, supports scientists in the Plan element of the Science Data Lifecycle Model. A basic USGS template has been created in the DMPTool to facilitate uptake by the agency's science staff.

Online Metadata Editor: https://www1.usgs.gov/csas/ome
The Online Metadata Editor (OME) is an online metadata creation tool developed by USGS that allows USGS and Department of the Interior scientists to document their data using the

JeSLIB 2017; 6(2): e1117
doi:10.7191/jeslib.2017.1117

Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata (FGDC-CSDGM) by answering simple questions about their data. This tool supports the Document element of the Model.

Metadata Wizard: https://www.sciencebase.gov/catalog/item/50ed7aa4e4b0438b00db080a
The Metadata Wizard tool, an ArcCatalog plugin developed by USGS, supports geospatial data users with creation and editing of metadata compliant with the FGDC-CSDGM standard. This tool also supports the Document element of the Model.

DOITool: https://www1.usgs.gov/csas/doi/index.htm
The Digital Object Identifier Tool, developed by USGS, allows creation of a permanent DOI for released USGS science data. This tool facilitates minting of digital object identifiers from DataCite, a prominent organization hosting identifiers, and supports the Preserve and Publish/Share aspects of the Model. There have been over 2,000 digital object identifiers created thus far for science data using the DOITool.

ScienceBase: https://www.sciencebase.gov/catalog
ScienceBase is a data and information management infrastructure that enables data upload, documentation, sharing, and dynamic data services using standards-compliant methods and technological components. ScienceBase, developed by USGS, furnishes a foundation for data stewardship, government open data, and scientific discovery. It serves as a data release platform and repository for USGS data, representing the Publish/Share in addition to Backup & Secure, and Preserve aspects of the Model. Over 900 scientists have released USGS data in ScienceBase over a one year period.

USGS Science Data Catalog: http://data.usgs.gov/datacatalog
The USGS Science Data Catalog (SDC) serves as a single access point for public USGS scientific datasets and as a provider to external catalogs in response to Federal Open Data requirements. The SDC allows the public to access USGS datasets through text and GIS-based search; topical browse; and keyword, data source, and scientist author name faceting. USGS policy requires metadata records that describe datasets be registered in the SDC as part of the Publish/Share element of the Model. The Catalog contains over 10,000 records available for public access and use.

Conclusion

The U.S. Geological Survey produces large-scale, integrated, world-class science through investigations that result in impartial information products and science data that are utilized by a wide audience. USGS recognizes a critical need to support its scientists in data management to ensure that science data are made available, understood, preserved, and reused. The agency developed a Science Data Lifecycle Model to frame data management policies, tools, education, and instruction. The Science Data Lifecycle Model became the basis for new, sweeping data management policies and the tools for policy implementation. The USGS policies underscore the importance of understanding the lifecycle science records undertake, the role data management plans play, the criticality of good metadata, and the responsibility to ensure science data are preserved for future use. The U.S. National Oceanic and Atmospheric Administration accurately sums up the responsibility of Federal research agencies by stating, “Accurate, timely, and comprehensive observations of the Earth and its surrounding space are

critical to support government decisions and policies, scientific research, and the economic, environmental, and public health of the nation and the world." The USGS Science Data Lifecycle Model helps address our responsibilities as it has become an influential and consistent way to express the various activities required for successful data management, and emphasizes the USGS' commitment to caring for essential and significant Earth science data. The Model is expected to serve the USGS for many years. It will continue to be the basis for all future data management policies, and serve as a framework for science data management applications.

Supplemental Content

An online supplement to this article can be found at http://dx.doi.org/10.7191/jeslib.2017.1117 under "Additional Files".

Disclosure

The authors report no conflict of interest.

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

Burwell, Sylvia M., Steven VanRoekel, Todd Park, Dominic J. Mancini. 2013. "Open Data Policy — Managing Information as an Asset." *Executive Office of the President, Office of Management and Budget* M-13-13. https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2013/m-13-13.pdf

Faundeen, John L., Thomas E. Burley, Jennifer A. Carlino, David L. Govoni, Heather S. Henkel, Sally L. Holl, Vivian B. Hutchison, Elizabeth Martín, Ellyn T. Montgomery, Cassandra C. Ladino, Steven Tessler, and Lisa S. Zolly. 2013. "The United States Geological Survey Science Data Lifecycle Model." *U.S. Geological Survey Open-File Report* 2013-1265. http://dx.doi.org/10.3133/ofr20131265

Holdren, John P. 2013. "Increasing Access to the Results of Federally Funded Scientific Research." *Executive Office of the President, Office of Science and Technology Policy*. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Beaujardière, Jeff De La. 2016. "NOAA Environmental Data Management." *Journal of Map & Geography Libraries* 12(1): 5-27. http://dx.doi.org/10.1080/15420353.2015.1087446

Obama, Barack. 2013. "Executive Order — Making Open and Machine Readable the New Default for Government Information." *Federal Register* 48(93). https://obamawhitehouse.archives.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-

U.S. Geological Survey. 2017a. "Fundamental Science Practices: Foundation Policy." *U.S. Geological Survey Manual* 502.01. Last modified August 30. https://www2.usgs.gov/usgs-manual/500/502-1.html

———. 2017b. "Fundamental Science Practices: Review, Approval, and Release of Information Products." *U.S. Geological Survey Manual* 502.04. Last modified August 30. https://www2.usgs.gov/usgs-manual/500/502-4.html

———. 2017c. "Fundamental Science Practices: Metadata for USGS Scientific Information Products Including Data." *U.S. Geological Survey Manual* 502.07. Last modified August 30. https://www2.usgs.gov/usgs-manual/500/502-7.html

JeSLIB 2017; 6(2): e1117
doi:10.7191/jeslib.2017.1117

———. 2017d. “Fundamental Science Practices: Preservation Requirements for Digital Scientific Data.” *U.S. Geological Survey Manual* 502.09. Last modified August 30. https://www2.usgs.gov/usgs-manual/500/502-9.html

———. 2017e. *Data Management Website*. Last Modified September 26. http://www2.usgs.gov/datamanagement/index.php