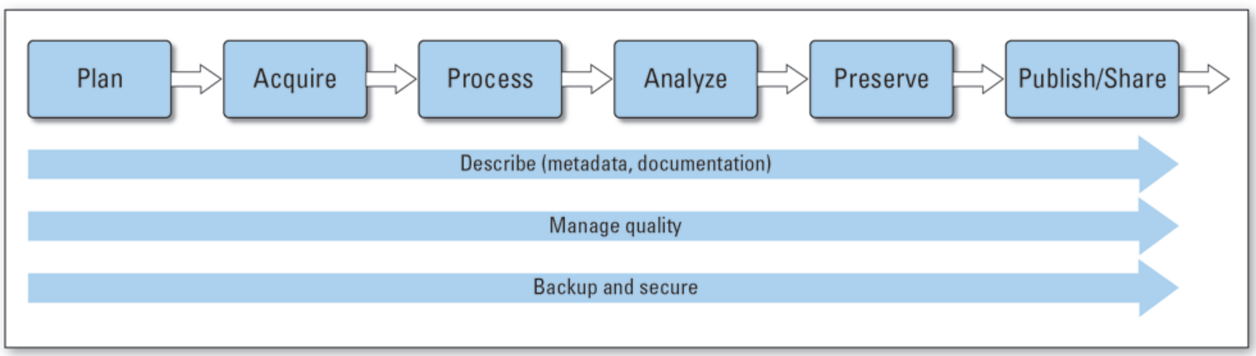
The United States Geological Survey Science Data Lifecycle Model



Open-File Report 2013–1265

U.S. Department of the Interior
U.S. Geological Survey

Cover. Diagram showing USGS Science Data Lifecycle Model.

The United States Geological Survey Science Data Lifecycle Model

By John L. Faundeen, 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

Open-File Report 2013–1265

U.S. Department of the Interior
U.S. Geological Survey

U.S. Department of the Interior
SALLY JEWELL, Secretary

U.S. Geological Survey
Suzette M. Kimball, Acting Director

U.S. Geological Survey, Reston, Virginia: 2013

For more information on the USGS—the Federal source for science about the Earth, its natural and living resources, natural hazards, and the environment, visit http://www.usgs.gov or call 1–888–ASK–USGS.

For an overview of USGS information products, including maps, imagery, and publications, visit http://www.usgs.gov/pubprod

To order this and other USGS information products, visit http://store.usgs.gov

Suggested citation:
Faundeen, J.L., Burley, T.E., Carlino, J.A., Govoni, D.L., Henkel, H.S., Holl, S.L., Hutchison, V.B., Martín, Elizabeth, Montgomery, E.T., Ladino, C.C., Tessler, Steven, and Zolly, L.S., 2013, The United States Geological Survey Science Data Lifecycle Model: U.S. Geological Survey Open-File Report 2013–1265, 4 p., *http://dx.doi.org/10.3133/ofr20131265*.

Contents

Figure

The United States Geological Survey Science Data Lifecycle Model

By John L. Faundeen, 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

Abstract

U.S. Geological Survey (USGS) data represent corporate assets with potential value beyond any immediate research use, and therefore need to be accounted for and properly managed throughout their lifecycle. Recognizing these motives, a USGS team developed a Science Data Lifecycle Model (SDLM) as a high-level view of data—from conception through preservation and sharing—to illustrate how data management activities relate to project workflows, and to assist with understanding the expectations of proper data management. In applying the Model to research activities, USGS scientists can ensure that data products will be well-described, preserved, accessible, and fit for re-use. The Model also serves as a structure to help the USGS evaluate and improve policies and practices for managing scientific data, and to identify areas in which new tools and standards are needed.

Background

The U.S. Geological Survey (USGS) Community for Data Integration (CDI) was established in 2009 (see *https://my.usgs.gov/confluence/display/cdi/CDI+Charter*) to address data and information management issues affecting the Bureau's scientific research. The CDI brings together expertise from external partners and representatives across USGS who are involved in research, data management, and information technology, and provides a forum for collaboration and brainstorming. Through partnerships and working groups, the CDI leads the development of data management tools and practices, cyber infrastructure, collaboration tools, and training in support of scientists and technology specialists. The CDI represents a dynamic community of practice focused on advancing science data and information management and integration capabilities across the USGS.

In 2010, the CDI established a Data Management Working Group (DMWG) to develop, enhance, and recommend best practices and policies that would assist USGS in effectively handling, documenting, preserving, and providing access to the Bureau's science data. To provide a conceptual foundation for these activities that encapsulates the best practices of data management in a model that resonates with USGS research scientists, a working group sub-team investigated existing data lifecycle models. That sub-team investigation resulted in a compilation of more than 50 data lifecycle models from academia, other Federal agencies, and international organizations. The models differed widely in granularity, complexity, presentation, and community perspective; however, none of the reviewed models captured traditional USGS processes or the level of detail and clarity desired. A simplified, custom model, informed by the reviewed models, was needed to clearly connect data management activities with research project plans and to provide a useful structure for organizing data management resources.

The sub-team analyzed each model for its relevance to existing USGS processes for creation and management of science data. Common characteristics among all models were identified, and the resulting information was synthesized and analyzed over several months by a review team with diverse perspectives on USGS science data; these efforts culminated in a two-day workshop to develop a first rendition of a model. Circular, spiral, linear, and decision-tree type models were posited and thoroughly discussed for their relevance to USGS processes and their effectiveness in communicating the stages of the data lifecycle.

The sub-team ultimately chose a linear presentation for the model graphic (fig. 1, hereafter referred to as the Model) having an intuitive, left-to-right flow of its narrative and a clearly defined starting point that aligns with the inception of the research project. Careful attention was paid to making effective semantic choices, with the firm goal of providing clear communication of the Model's components. Explanatory text accompanies the Model and is intended to help users better understand each element. The Model's supplementary documentation and discussion are available on the USGS Data Management Web site (*http://www.usgs.gov/datamanagement/*).

The sub-team made drafts of the Model available for review through multiple venues, including hosting a poster at the 2011 CDI Workshop, briefings to the CDI during monthly

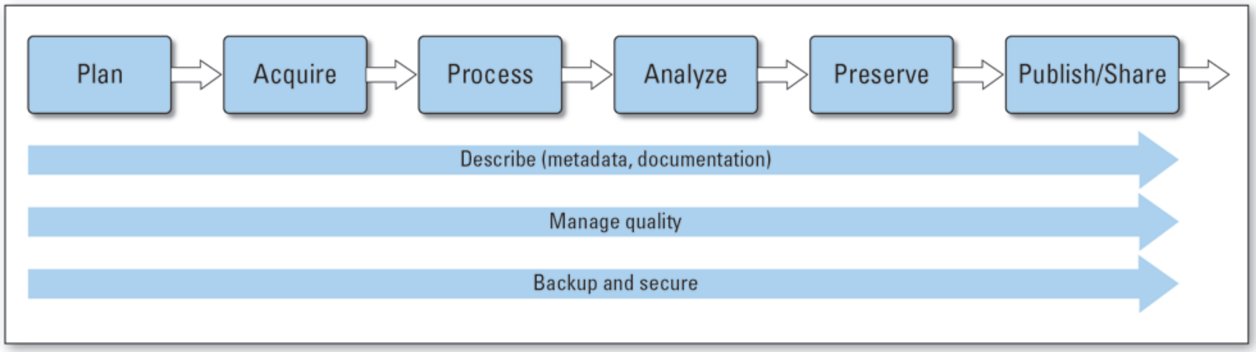


Figure 1. USGS Science Data Lifecycle Model. Boxes indicate the main Model elements, and the shaded arrows below represent cross-cutting elements.

meetings, and posting on the CDI Web site (*https://my.usgs.gov/confluence/display/cdi/Home*). USGS scientists, data managers, and policy analysts provided valuable feedback and suggested changes that were incorporated into subsequent iterations. The final draft of the Model was reviewed and accepted in November 2012 by CDI executive sponsors Kevin T. Gallagher (Associate Director for Core Science Systems) and Linda C. Gundersen (Emeritus; then-Director, Office of Science Quality and Integrity).

The Data Lifecycle Model

The USGS serves the Nation by providing scientific information; therefore, data are a core asset to the Bureau. The February 22, 2013, Memorandum from the White House Office of Science and Technology Policy (Holdren, 2013) observes that "Scientific research supported by the Federal Government catalyzes innovative breakthroughs that drive our economy. The results of that research become the grist for new insights and are assets for progress in areas such as health, energy, the environment, agriculture, and national security."

The Bureau's Fundamental Science Practices encourage "USGS scientists…to publish their data and findings in ways that contribute to the most effective release of USGS science and best enhance the Bureau's reputation for reliable science" (U.S. Geological Survey, 2011). Proper management of USGS science data is the responsibility of USGS scientists and the staff who support them. Data management and data integration have been identified in the USGS Science Strategy as critical areas that are essential to the success of future understanding and application of USGS science (U.S. Geological Survey, 2007). Implementing the Model helps USGS scientists understand and manage for the lifecycle of data and information products to ensure that the data are discoverable, well described, and preserved for access and use beyond the life of research projects.

Incorporating data management principles and practices is an essential component of the larger process of planning activities, resources, and techniques that will be required by the research project. A data lifecycle model offers a high-level overview of the individual actions, operations, or processes that must be undertaken at different stages. The Model as a visual tool can, in turn, assist scientists in anticipating and planning for specific actions that need to be taken at each stage to manage the data, and thus help to ensure timely, comprehensive, and secure approaches to data curation. The resulting well-curated data resources, which researchers can re-use, are critical to integrated science and extend the value of the data. This Model, developed for USGS science data, is one way to facilitate shared recognition and understanding of the necessary steps to document, protect, and make available the Bureau's valued data assets.

Primary Model Elements

Plan, the first Model element, is intended to assist scientists in assuring consideration of all activities related to the handling of the project's data assets, from project inception to publication and archiving. During this stage, all elements of the Model should be evaluated, addressed, and documented. The project team should consider approaches, needed resources (including funding and personnel), and intended outputs for each stage of the data lifecycle. A data management plan is the recommended output of this element of the Model.

Acquire, the second Model element, represents the activities through which new or existing data are collected, generated, or considered and evaluated for re-use. Streamgage data, historical maps, seismology motion sensor outputs, biological records, and satellite observations are examples of acquired data and information that represent the diverse and robust variety of science data inputs to USGS research. Scientists are skillful in designing data acquisition techniques to address research questions; in the USGS context, this element emphasizes the importance of considering relevant USGS policies and best practices that maintain the provenance and integrity of the data as a USGS information product. The outputs of this element are the project's data inputs.

Process, the third Model element, represents various activities associated with preparation of new or previously collected data inputs. Processing of input data may entail definition of data elements; integration of disparate datasets; extraction, transformation, and load operations; and application of calibrations to prepare the data for analysis. The Process element in the Model reminds scientists that USGS standards and tools are available that can meet project requirements while also building a Bureau-wide foundation of data for integrated science. The outputs of this element are datasets that are ready for integration and analysis.

Analyze, the fourth Model element, represents the activities associated with the exploration and interpretation of processed data, where hypotheses are tested, discoveries are made, and conclusions are drawn. Analytical activities include summarization, graphing, statistical analysis, spatial analysis, and modeling, and are used to produce scientific results and information. In this element new data are generated, versions are tracked, and processes are documented. Data management during analysis improves the efficiency of data analysis activities, preserves documentation that is critical for scientific integrity, and creates a foundation for future research. The outputs of this element are interpretations or new datasets, which often are published in written reports or machine-readable formats such as map layers or numerical modeling results.

Preserve, the fifth Model element, represents the activities associated with storing data for long-term use and accessibility. Preservation often is not considered until the end stage of a project, when it might be neglected because of the pressure of project budgets and timetables. The intentional placement of this element ahead of Publish/Share in the Model is a reminder that federally funded scientists must plan for the long-term preservation of data, metadata, ancillary products, application-neutral storage formats, and any additional documentation, to ensure availability and re-use.

Publish/Share, the sixth element in the Model, combines the Bureau's concepts of traditional peer-reviewed publication with the distribution of data through Web sites, data catalogs, social media, and other venues. Publication and dissemination of USGS data and information are critical components of the USGS Mission (U.S. Geological Survey, 1989), and are a focal point of recent Federal directives to increase access to the results of federally funded research (Holdren, 2013). This element reminds scientists that data, as well as traditional publications, are research products.

Cross-Cutting Model Elements

Each of the primary elements of the Model addresses discrete activities and outputs unique to that stage; however, other critical activities must be performed continually across all stages of the lifecycle to help support effective data management (fig. 1).

Describe (metadata, documentation), the first crosscutting element, highlights the importance of step-wise documentation throughout the data lifecycle. Beginning with the data management plan, this element emphasizes documentation of every lifecycle stage in sufficient detail that other scientists can validate research outputs through replication, evaluate the validity of the results, and determine the usefulness of the data for future research. Recording this information at each stage of the data lifecycle—rather than at the concluding stages of the project—helps to ensure accuracy and comprehension of the science data created, compiled, processed, and shared. Standards-based metadata and documentation such as software code comments, data models, and work flows facilitate indexing, accession, understanding, and future uses of the data. Although misinterpretation or misuse of data cannot be prevented, well-documented data can help to expose and correct such errors when they occur. Finally, the Describe element communicates the provenance and authority of USGS data and information.

Manage quality, the second crosscutting element, reminds scientists to plan quality-assurance measures for data at the project's inception, and then undertake ongoing quality-control monitoring and adjustment at subsequent lifecycle stages to verify that those measures perform as expected as the project proceeds. This element aligns with the Bureau's emphasis on releasing only products of the highest quality science.

Backup and secure, the third crosscutting element, involves managing physical risks to the data throughout the data lifecycle while also ensuring that the data are accessible. This element reminds scientists that routine backups are critical to prevent the physical loss of data because of hardware or software failure, natural disasters, or human error before the final Preservation of the data. Loss-prevention measures apply to the raw and processed research data, original science plan, data management plan, data acquisition strategy, processing procedures, versioning, analysis methods, published products, and associated metadata. This element also encourages scientists to plan for secure data sharing services, particularly when the project's scientists work at multiple facilities.

Data Management Roles and Responsibilities in Research Projects

The USGS Science Data Lifecycle Model includes data management activities that require specialized knowledge and skills, as well as ongoing education about methods and standards. The Model encourages researchers to plan project teams that recognize two different roles: researchers and data stewards. Depending on staff expertise, the researcher and data steward roles may fall to one person for one or more lifecycle stages or may be divided among multiple individuals with discrete responsibilities. Although multiple personnel may oversee the various data lifecycle elements, the project manager is responsible for ensuring that each element is addressed throughout the life of the project.

Summary

The U.S. Geological Survey (USGS) Science Data Lifecycle Model represents a high-level view of USGS data collection, handling, and dissemination activities. In applying the Model to research activities, USGS scientists can ensure that data products will be well described, preserved, accessible, and fit for re-use. The Model also serves as a structure to help evaluate and improve USGS policies and practices for managing scientific data, and to identify areas in which new tools and standards are needed.

Acknowledgments

This work benefited from USGS staff contributions provided by Richard Huffine, Tim Kern, Qi Tong, and USGS Volunteer Trent Faust, at various times through the development of the Model. The authors recognize and thank them for their efforts.

"Government information shall be managed as an asset throughout its lifecycle to promote interoperability and openness, and, wherever possible and legally permissible, to ensure that data are released to the public in ways that make the data easy to find, accessible, and usable." — Executive Order — Making Open and Machine Readable the New Default for Government Information (May 9, 2013)

References Cited

Executive Order No. 13642, Federal Register Vol. 48, No. 93, pg. 28111 (May 14, 2013), accessed September 2013 at *https://www.federalregister.gov/articles/2013/05/14/2013-11533/making-open-and-machine-readable-the-new-default-for-government-information#p-4*.

Holdren, J.P., 2013, Memorandum for the heads of executive departments and agencies—Increasing access to the results of federally funded scientific research: Executive Office of the President, Office of Science and Technology Policy, 6 p., accessed February 26, 2013, at *http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf*.

U.S. Geological Survey, 1989, U.S. Geological Survey Manual—Organization—Creation, Mission, and Functions: U.S. Geological Survey Manual Section 120.1, accessed September 2012 at *http://www.usgs.gov/usgs-manual/120/120-1.html*.

U.S. Geological Survey, 2007, Facing tomorrow's challenges—U.S. Geological Survey science in the decade 2007–2017: U.S. Geological Survey Circular 1309, x + 70 p. Also available at *http://pubs.usgs.gov/circ/2007/1309/*.

U.S. Geological Survey, 2011, U.S. Geological Survey Manual Chapter 502.4—Fundamental Science Practices—Review, Approval, and Release of Information Products, accessed September 2013 at *http://www.usgs.gov/usgs-manual/500/502-4.html*.

Publishing support provided by:
Rolla Publishing Service Center

For more information concerning this publication, contact:
U.S. Geological Survey Earth Resources Observation
and Science (EROS) Center
47914 252nd Street
Sioux Falls, South Dakota 57198
(605) 594-6151

Or visit the EROS Center Web site at:
http://eros.usgs.gov/

Faundeen and others—**The United States Geological Survey Science Data Lifecycle Model**—Open-File Report 2013–1265

ISSN 2331–1258 (online)
http://dx.doi.org/10.3133/ofr20131265