

Science Synthesis, Analysis, and Research Program

USGS Enterprise Tools for Efficient and Effective Management of Science Data

The Science Data Management Branch (SDM) of the U.S. Geological Survey (USGS) provides data management expertise and leadership and develops guidance and tools to support the USGS in providing the nation with reliable scientific information on the basis of which to describe the Earth. The SDM suite of tools supports the USGS Data Management Lifecycle by facilitating quality assurance, description, curation, and publishing of the Bureau's scientific data. The SDM suite of tools includes the USGS Data Management Website, USGS Science Data Catalog, Digital Object Identifier Tool, ScienceBase, ScienceBase Data Release Tool, Metadata Wizard, and Online Metadata Editor.

Introduction

The Science Data Management Branch (SDM) of the U.S. Geological Survey (USGS) provides data management expertise and leadership and develops guidance and tools to support the USGS in providing the nation with reliable scientific information to describe and understand the Earth. SDM is part of the Science Analytics and Synthesis (SAS) organization in the USGS Core Science Systems Mission Area. SDM works to facilitate and implement each stage of the USGS Data Lifecycle (Faundeen and others, 2013) (fig. 1). SDM also adds value by enabling science data reuse and integration and assists scientists in adhering to the expectations of both the USGS Fundamental Science Practices (FSP) (USGS, 2006), and the USGS Public Access Plan (USGS, 2016).

The USGS has long recognized and followed a commitment to data management best practices that are key to producing quality scientific data and making them available to the public. Multiple broad Federal Government policies specific to open, accessible, and discoverable scientific data (Holdren, 2013), coupled with a global emphasis on FAIR (Findable, Accessible, Interoperable, and Reusable) data principles (Wilkinson and

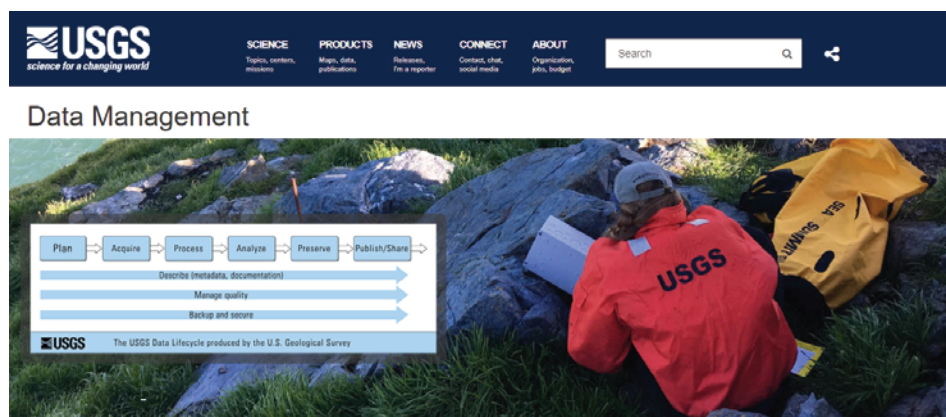


Figure 2. USGS Data Management website home page.

others, 2016), reinforce a need for enterprise Bureau applications such as those described in this fact sheet.

Enterprise Tools to Manage Scientific Data in USGS

The SDM's suite of tools facilitates quality assurance, description, curation, and publishing of the Bureau's scientific data. Each tool relates directly to components of the USGS Data Lifecycle and is flexible enough to accommodate various data management workflows. The tools include the USGS Data Management Website, USGS Science Data Catalog, Digital Object Identifier Tool, ScienceBase,

ScienceBase Data Release Tool, Metadata Wizard, and Online Metadata Editor. SDM designed these tools to be connected by Application Programming Interfaces (APIs), and thus to create more efficient workflows. The SDM tool suite can also facilitate data usage and citation tracking as one measure of the effects of the data in the scientific community.

USGS Data Management Website

The USGS Data Management Website (<https://www.usgs.gov/products/data-and-tools/data-management>) (fig. 2) introduces users to data management concepts and provides guidance regarding best practices and tools for each stage of the USGS Data Lifecycle. The Data Lifecycle (fig. 1) serves as the organizational framework for the Website's content. The Website also provides information that helps researchers meet the data management requirements of FSP. In 2013, the Website was awarded the USGS Shoemaker Award for Communications Excellence, the USGS's most prestigious award for communications.

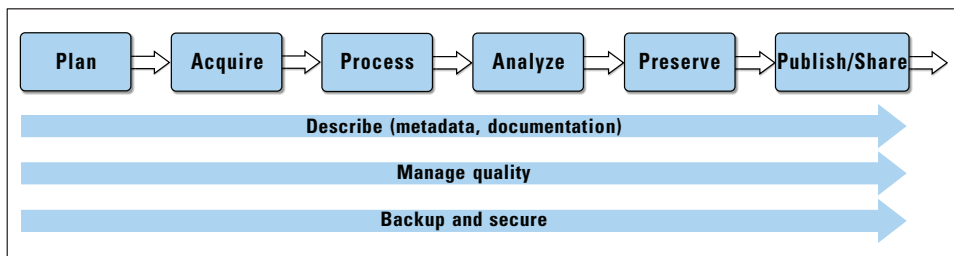


Figure 1. USGS Data Lifecycle. The USGS Data Lifecycle aligns data-management activities with project workflows and guides proper data management.

USGS Science Data Catalog (SDC)

The USGS Science Data Catalog (SDC) (<https://data.usgs.gov/datacatalog>) (fig. 3) is a comprehensive metadata catalog that facilitates discovery of and access to USGS research and monitoring data. Users can use SDC to interact with thousands of publicly accessible USGS scientific data records by using text search and browse features or a map interface. SDC automatically harvests structured metadata from ScienceBase and other USGS web-accessible locations and publishes this content in other catalogs, including the Department of the Interior Catalog (<https://data.doi.gov/dataset>) and [data.gov](https://www.data.gov) (<https://www.data.gov>), for broader distribution. SDC also serves as a reporting mechanism for the USGS to the Office of Management and Budget, a requirement of the USGS Public Access Plan. As of April 2020, the catalog included more than 21,000 metadata records describing published USGS data assets.

USGS Digital Object Identifier Tool

The USGS Digital Object Identifier (DOI) Tool (<https://www1.usgs.gov/csas/doi/#/>) (fig. 4) allows USGS personnel to generate globally unique, persistent, and resolvable digital object identifiers (DOIs) for their online scientific products. DOIs provide a permanent Universal Resource Locator (URL) to the data while allowing the actual physical path to the online product to be updated by the data owners over time as files move to new server locations. When managed appropriately, DOIs help ensure continuous access to the data over time and facilitate product usage and citation tracking after release. To the public, DOIs

Figure 4. USGS Digital Object Identifier Tool home page.

are an important mechanism for persistently locating scientific data for understanding and potential reuse.

DOIs are reserved by interacting directly with the DOI Tool's user interface or programmatically through the API. Following USGS FSP, DOIs can be assigned to products prior to formal release. Once all approvals are final, the DOI is made public and published to DataCite (<https://datacite.org/>), a global non-profit organization that provides a registry and handle service for research data DOIs.

ScienceBase

ScienceBase (<https://www.sciencebase.gov/catalog/>) (fig. 5) is a USGS Trusted Digital Repository (USGS, 2020) and a primary host location in which Bureau scientists can store and publish data. ScienceBase is developed and maintained by SDM and provides a collaborative data management platform on which scientists and data stewards can document, organize,

preserve, and publish scientific data products and other science-related resources. ScienceBase provides well-organized documented datasets and scientific information through a browser, web services, and an API, adding value to digital science data. ScienceBase facilitates data management workflows by providing direct connections and transfer of information among other USGS systems and tools, such as the USGS Science Data Catalog (described above). Since 2015, SDM has continued to refine standard operating procedures for releasing USGS data products through ScienceBase and helps USGS researchers comply with the FSP policies (USGS, 2017). This process, initiated through the ScienceBase Data Release (SBDR) Tool, optimizes the sequence of actions needed to complete a USGS data release. ScienceBase also supports tools and workflows of other groups, providing them with efficient web-based search functions and machine accessibility.

ScienceBase Data Release Tool (SBDR)

The SBDR Tool (fig. 6) streamlines the formal USGS data release process in ScienceBase. SBDR is a web application that creates an individual ScienceBase landing page for each data release in the repository. Additionally, SBDR reserves a DOI for each release by seamlessly accessing the DOI Tool. Similarly, through a connection with the USGS Information Product Tracking System, an internal review and approval

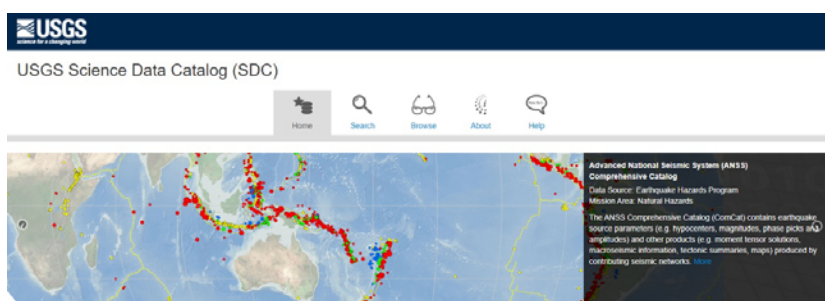


Figure 3. USGS Science Data Catalog home page.

tracking application, SBDR automatically reduces data entry by retrieving existing information about a data release. SBDR can be configured to notify specified data managers that a new data release has been initiated. The SDM team can also configure the SBDR for each USGS Science Center by allowing in-house data managers to control permissions to read and edit repository landing pages, manage the associated DOI(s), and notify the appropriate staff of newly initiated data releases. SBDR can also identify metadata records for automatic transmission to the SDC upon final approval of a data release and offer detailed metrics on the number of data releases produced by a USGS Science Center or individual author.

Metadata Wizard 2.0

The USGS Metadata Wizard (fig. 7) simplifies metadata creation by extracting information directly from geospatial datasets where possible, such as coordinate reference info, spatial extent, feature counts, entity and attribute information, and a jargon-free interface to collect other relevant information about the data. Metadata Wizard was originally designed for direct integration with ArcGIS Desktop to help users create Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM) compliant metadata. The redesigned Metadata Wizard 2.0 offers continued



Figure 5. USGS ScienceBase home page.

Please fill out and submit this form to start a data release in ScienceBase. A new landing page with a digital object identifier (DOI) will be created. You will receive an email with a link to the new landing page and further instructions on how to complete the data release. Visit the [ScienceBase Data Release Instructions](#) page to review the ScienceBase data release workflow and frequently asked questions.

* required entry

*** IPDS number**
To get an IPDS number for your data release go to <https://ipds.usgs.gov>, create a new record, and select "Data Release" as the product type (e.g. IP-012345). To use the autofill feature, please connect to the USGS network.

IP: [Autofill Form with IPDS Information](#)

*** Data Release Title**
Please choose a title that is descriptive. This title will be used in the data release citation and will be discoverable in ScienceBase search results.

e.g. Boat-based water-surface elevation surveys along the upper Willamette River, Oregon, in March 2015

*** USGS Mission Area**
Please select the Mission Area with which this data release is associated.

*** Science Center**
Please select the science center or program that has primary responsibility for this data release.

Figure 6. ScienceBase data release tool.

support for ArcGIS users, but also can be run as a stand-alone desktop application, compatible with both Macintosh and Windows operating systems. The Metadata Wizard can be integrated with several data management tools that help users complete their metadata records. Tools include the USGS Metadata Parser (<https://mrdata.usgs.gov/validation/>) to validate metadata structure; USGS Thesaurus (<https://www2.usgs.gov/science/about/>), a controlled vocabulary of keywords named the Integrated Taxonomic Information System (ITIS) (<https://www.itis.gov/>) – an authoritative biological taxonomic reference; and DataCite and CrossRef (<https://www.crossref.org/>) – sources of bibliographic information for related publications. USGS-employee contact information can also be easily pulled into the tool. Metadata Wizard can also

Figure 7. Metadata Wizard 2.0 home page.

connect directly to an Extensible Markup Language (XML) metadata file associated with a landing page in ScienceBase that allows data managers to directly edit existing metadata files. Metadata Wizard 2.0 is bundled with a version of Anaconda (<https://www.anaconda.com>), a data science platform, which provides data managers a simplified way to begin batch-processing of metadata files and complete other data management tasks with the programming language Python (™) (<https://www.python.org>) and the open-source code sharing applications, Jupyter Notebooks (™) (<https://jupyter.org>). Metadata Wizard and Metadata Wizard 2.0 were designed and built in collaboration between the SAS SDM Branch and the USGS Fort Collins Science Center.

Online Metadata Editor (OME)

The Online Metadata Editor (OME) (fig. 8) is a web application to help USGS scientists describe their data or product by asking simple, jargon-free questions. OME

generates a FGDC CSDGM-compliant record (FGDC, 2020) in XML, a standard metadata-record format commonly used by the USGS. Users can log in to OME and create new metadata records or upload and edit existing records, view a list of all the metadata records they have created or uploaded in the past, and save partial metadata records to complete later. OME is connected to other USGS tools, such as ITIS, USGS Metadata Parser, DOI Tool, and an internal USGS Employee Directory to aid in metadata description and production. Metadata records created by using OME can populate a ScienceBase repository page for a data release and can also be submitted to the USGS SDC.

References Cited

Faundeen, J.L., Burley, T.E., Carlino, J.A., Govoni, D.L., Henkel, H.S., Holl, S.L., Hutchison, V.B., Martin, E., Montgomery, E.T., Ladino, C.C., Tessler, S., 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., accessed February 6, 2020 at <https://doi.org/10.3133/ofr20131265>.

Federal Geographic Data Committee (FGDC), 2020, Content Standard for Digital Geospatial Metadata (CSDGM), Ver. 2 (FGDC–STD–001198), accessed February 7, 2020, at <https://www.fgdc.gov/metadata/csdgm-standard>.

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 (February 22, 2013), 6 p., accessed February 6, 2020 at https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.

U.S. Geological Survey [USGS], 2006, Fundamental Science Practices (FSP): accessed January 16, 2020, at <https://www.usgs.gov/about/organization/science-support/science-quality-and-integrity/fundamental-science-practices>.

U.S. Geological Survey [USGS], 2017, 502.8—Fundamental Science Practices—Review and approval of scientific data for release: U.S. Geological Survey, accessed January 16, 2019, at <https://www.usgs.gov/about/organization/science-support/survey-manual/5028-fundamental-science-practices-review-and>.

U.S. Geological Survey [USGS], 2020—USGS Trusted Digital Repositories (TDR), accessed February 7, 2020, at <https://www.usgs.gov/about/organization/science-support/office-science-quality-and-integrity/trusted-digital-repository>.

U.S. Geological Survey [USGS], 2016, Public access to results of federally funded research at the U.S. Geological Survey—Scholarly publications and digital data: U.S. Geological Survey, accessed January 8, 2016, at <https://prd-wret.s3.us-west-2.amazonaws.com/assets/palladium/production/atoms/files/USGS-PublicAccessPlan-APPROVED-v1-03.pdf>.

Wilkinson, M., Dumontier, M., Aalbersberg, I., and others, 2016, The FAIR guiding principles for scientific data management and stewardship: Scientific Data v. 3, no. 160018, 9 p., accessed February 12, 2020 at <https://doi.org/10.1038/sdata.2016.18>.

Figure 8. Online Metadata Editor home page.

By Vivian B. Hutchison, Amanda Liford, Ricardo McClees-Funinan, Lisa Zolly, Drew Ignizio, Madison Langseth, Brandon Serna, Elizabeth Sellers, Leslie Hsu, Tamar Norkin, Marcia McNiff, and Grace Donovan
For more information contact Vivian Hutchison, Branch Chief, USGS SAS Science Data Management (vhutchison@usgs.gov)