

Leveraging Existing Technology: Developing a Trusted Digital Repository for the U.S. Geological Survey

Vivian B. Hutchison
U.S. Geological Survey

Tamar Norkin
U.S. Geological Survey

Madison L. Langseth
U.S. Geological Survey

Drew A. Ignizio
U.S. Geological Survey

Lisa S. Zolly
U.S. Geological Survey

Ricardo McClees-Funinan
U.S. Geological Survey

Amanda N. Liford
U.S. Geological Survey

Abstract

As Federal Government agencies in the United States pivot to increase access to scientific data (Sheehan, 2016), the U.S. Geological Survey (USGS) has made substantial progress (Kriesberg et al., 2017). USGS authors are required to make federally funded data publicly available in an approved data repository (USGS, 2016b). This type of public data product, known as a USGS data release, serves as a method for publishing reviewed and approved data. In this paper, we present major milestones in the approach the USGS took to transition an existing technology platform to a Trusted Digital Repository. We describe both the technical and the non-technical actions that contributed to a successful outcome. We highlight how initial workflows revealed patterns that were later automated, and the ways in which assessments and user feedback influenced design and implementation. The paper concludes with lessons learned, such as the importance of a community of practice, application programming interface (API)-driven technologies, iterative development, and user-centered design. This paper is intended to offer a potential roadmap for organizations pursuing similar goals.

Received 23 September 2020 ~ *Revision received* 11 April 2021 ~ *Accepted* 11 April 2021

Correspondence should be addressed to Vivian B Hutchison, US Geological Survey, Denver Federal Center, West 6th Avenue & Kipling, Building 810, MS302, Lakewood, CO 80225 Email: vhutchison@usgs.gov

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: <http://www.ijdc.net/>

Copyright rests with the authors. This work is released under a Creative Commons Attribution Licence, version 4.0. For details please see <https://creativecommons.org/licenses/by/4.0/>



Introduction

Open access to scientific data is an emerging direction that governments and research institutions around the world are embracing (Arzberger et al., 2004; Janssen et al., 2012; Molloy, 2011). Starting in 2013, the U.S. Federal Government has issued multiple directives to its agencies to make federally funded science data available and open to the public (Holdren, 2013; Burwell et al., 2013; Exec. Order No. 13642, 2013). The goals of the directives were to increase the return on public investment, ‘accelerate scientific breakthroughs and innovation, promote entrepreneurship, and enhance economic growth and job creation’ (Holdren, 2013). By following these directives and initiatives such as FAIR (Findable, Accessible, Interoperable, Reusable data) (Wilkinson et al., 2016), government science organizations are transitioning towards a more mature model for supporting public access to research outputs.

The U.S. Geological Survey (USGS), a Bureau within the U.S. Department of the Interior, is one of the Federal agencies going through this transition. As a government science agency, the USGS provides scientific research, knowledge, and data in the Earth sciences to a wide variety of stakeholders. It is organized in a dispersed structure, with science centers located in every state in the country. Its diverse mission covers Earth science topics including natural hazards, water, energy, minerals, ecosystems, and the effects of climate and land-use change.

Since its establishment in 1879, the USGS has been a responsible steward of the vast array of data its scientists collect and has a rich history of making scientific results available through peer-reviewed publications.

The 2016 USGS Public Access Plan, entitled ‘Public Access to Results of Federally Funded Research at the U.S. Geological Survey: Scholarly Publications and Digital Data’ (USGS, 2016a), was written in response to the 2013 Federal Government directives and provided a framework for how the USGS would respond to meet the requirements. The Public Access Plan built on existing USGS policies, the Fundamental Science Practices (FSP) (Fundamental Science Practices Advisory Committee, 2011) which provide requirements and best practices for how USGS scientists conduct research.

As part of the plan’s implementation, the USGS added a set of FSP policies (USGS, 2017a, 2017b, 2017c, 2017d) to address how scientists manage and release their data products (i.e., datasets representing scientific observations, measurements, or analytical outputs). The policies require a data management plan, standardized metadata that describe the data, formal review and approval of the data and metadata, a digital object identifier, storage within an approved repository, and submission of the metadata record(s) to the USGS Science Data Catalog¹. Responsibility for implementing these data policies was assigned to the individual USGS science centers, such that scientific data management activities remained with the expertise in the centers.

The implementation of the USGS Public Access Plan also required that the USGS evaluate current technologies and processes in order to leverage existing data systems when possible. The USGS identified a need for dedicated Bureau-approved repositories that scientists could rely on for public release of their data. This prompted a team of experts in the USGS to develop an internal approval process for certifying trusted digital repositories to support the release, curation, and preservation of USGS scientific data.

¹ USGS Science Data Catalog: <https://data.usgs.gov/datacatalog>

As part of this development, they adapted the general requirements of the Core Trust Seal criteria (Edmunds et al., 2016) to create a USGS-specific certification process. This evolving certification approach comprises a building block towards future USGS engagement in the international Core Trust Seal certification; in the initial implementation of the Public Access Plan, this approach was used to prompt existing data systems to consider the capabilities necessary to adhere to international standards for certification.

ScienceBase² was the first data storage and delivery platform approved through the USGS internal certification process to be a Trusted Digital Repository for the Bureau. First launched in 2009, ScienceBase was originally designed as a science data collaboration platform. It already featured capabilities that aligned with the requirements of the USGS Public Access Plan and was therefore well-positioned to meet the Bureau's needs for data publication. The USGS determined that the most cost-effective and efficient approach would be to adapt this existing application instead of creating a new repository. Updating ScienceBase to more specifically meet the need for data publication required a focused and iterative effort over several years. This paper describes the steps taken, capabilities developed, and lessons learned in the successful transition of ScienceBase from an existing data collaboration platform to a Trusted Digital Repository for the stewardship of USGS data. This paper offers a roadmap and considerations that may be relevant to other organizations planning similar projects.

Background

ScienceBase was designed as a web application for USGS scientists and their collaborators to store and share data, to manage digital collections with flexible permission controls, and to access resources both programmatically and through a web browser user interface. Prior to its approval as a public data release mechanism for USGS scientists, ScienceBase supported a variety of repository features including file upload and storage, permission-controlled access to content, the ability to generate and host web services for geospatial data, a robust application programming interface (API), and a publicly searchable user interface. ScienceBase uses a standardized data model with a consistent set of informational facets (e.g., title, abstract, keywords) to describe and organize stored content. The individual entries, or items, in ScienceBase are stored as JSON records, a common transfer standard that uses simple 'key': 'value' pairings for informational content. Every item supports granular permissions, enabling access controls by role, organizational grouping, or at the individual user level. The application is built on a MongoDB database integrated with Elasticsearch. This supports performant search and powerful querying capabilities that make it possible to integrate content from a large, multi-million record catalog into dynamic collections and to connect ScienceBase-hosted data with external applications and workflows. Search, read, and authenticated edit actions are exposed as REST API endpoints, which support integration with modern web workflows. Any user or tool can work with ScienceBase via a browser or the REST endpoints. Additionally, two libraries exist – one written in the R programming language³ and the other in the Python programming language⁴ – that facilitate easy use of ScienceBase directly from tools written in R or Python.

² ScienceBase: <https://www.sciencebase.gov>

³ ScienceBase R tools: <https://github.com/USGS-R/sbtools>

⁴ Python ScienceBase utilities: <https://github.com/usgs/sciencebasepy>

As the USGS considered options for meeting the USGS Public Access Plan requirements, leveraging ScienceBase to support a data release workflow was determined to be a natural next step. The ScienceBase application is managed by the Science Data Management (SDM) team within the USGS Science Analytics and Synthesis (SAS) program. During the transition timeframe, the SDM team consisted of ten employees with backgrounds and skills in data management, data science, and application development. Although multiple individuals contributed to various activities in ScienceBase, the core data release activity was supported by two full-time employees. During the time period of October 1, 2015 – October 1, 2020, the SDM team completed more than 3,500 data releases for 97 USGS science centers – a notable achievement considering the work hours available for process support.

The process of transitioning ScienceBase from a community collaboration platform to an established USGS Trusted Digital Repository (Latysh et al., 2020) required a number of strategic actions. These included an independent review of ScienceBase against criteria for well-managed data repositories (Stall et al., 2016); design of a user workflow to enable efficient ingest, organization, and display of data; and development of a mechanism to connect to other enterprise data management tools. The process has been iterative and has continued to evolve over time based on new considerations and user needs.

Development of a Data Release Workflow in ScienceBase

Starting in 2015 – during which new USGS policy requirements were drafted but not yet in effect – the SDM team enacted a series of steps to create a data release workflow in ScienceBase for USGS scientists. The first step was to study a set of use cases for USGS data publication. The SDM team wrote recommendations for each case and unified them to develop a single, flexible workflow for data authors, with an understanding that user needs and considerations might vary between different research projects. The resulting workflow was designed to provide a clear, step-by-step process for authors to follow and to help them meet all requirements of the USGS data release policies. Documentation created for users included online instructions, frequently asked questions (FAQs), tutorial videos, and a workflow diagram.

Figure 1 shows a standard data release landing page in ScienceBase. The main elements on the page – title, dates, citation, summary, contacts, and downloadable data files – are standardized and appear on all landing pages for data release products. Data authors have the option of including an image to display on the page and can add additional features such as a geographic footprint, spatial web services, and links to external resources.

ScienceBase Catalog → USGS Data Release Products → Shoreline change rates in sa...

Shoreline change rates in salt marsh units in Edwin B. Forsythe National Wildlife Refuge, New Jersey View

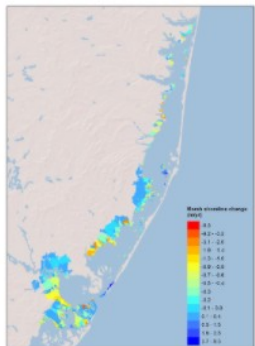
Dates
 Publication Date : 2018-01-11

Citation
 Defne, Zafer and Ganju, N.K., 2018, Shoreline change rates in salt marsh units in Edwin B. Forsythe National Wildlife Refuge, New Jersey: U.S. Geological Survey data release, <https://doi.org/10.5066/F7PN94K2>.

Summary
 Monitoring shoreline change is of interest in many coastal areas because it enables quantification of land loss over time. Evolution of shoreline position is determined by the balance between erosion and accretion along the coast. In the case of salt marshes, erosion along the water boundary causes a loss of ecosystem services, such as habitat provision, carbon storage, and wave attenuation. In terms of vulnerability, higher shoreline erosion rates indicate higher vulnerability.

This dataset displays shoreline change rates at the Edwin B. Forsythe National Wildlife Refuge (EBFNWR), which spans over Great Bay, Little Egg Harbor, and Barnegat Bay in New Jersey, USA. Shoreline change rates are based on Smith and Terrano (2017) analysis of digital vector shorelines acquired from historic topographic sheets, aerial photography, and/or lidar using the AMBUR package (Jackson, 2010). Linear Regression Rates (LRR) of shoreline change were averaged along the shoreline of each salt marsh unit to generate this dataset. Positive and negative values indicate accretion and erosion respectively.

As part of the Hurricane Sandy Science Plan, the U.S. Geological Survey is expanding National Assessment of Coastal Change Hazards and forecast products to coastal wetlands. The intent is to provide federal, state, and local [... show more ...](#)



Contacts
 Point of Contact : Zafer Defne
 Originator : Zafer Defne, Neil Kamal Ganju
 Publisher : U.S. Geological Survey
 Distributor : U.S. Geological Survey - ScienceBase
 USGS Mission Area : Natural Hazards
 SDC Data Owner : Woods Hole Coastal and Marine Science Center

Attached Files

File Name	Size
mu_LRR_EBFNWRp.shp.xml "CSDGM Metadata" Original FGDC Metadata	30.2 KB
mu_LRR_EBFNWRp.png "Browse graphic"	1.66 MB
mu_LRR_EBFNWRp.zip "Data and metadata for download"	18.74 MB

Spatial Services
 ArcGIS Mapping Service : <https://www.sciencebase.gov/arcg>
 WMS Service : <https://www.sciencebase.gov/arcg>

Communities
 • USGS Data Release Products
 • Woods Hole Coastal and Marine Science Center

Associated Items

Figure 1. Example of a data release landing page in ScienceBase (available at [doi:10.5066/F7PN94K2](https://doi.org/10.5066/F7PN94K2); Defne and Ganju, 2018).

Initial Workflow

In the first iteration of the ScienceBase data release workflow (Figure 2), the key steps in the process were completed manually by the SDM team. This was helpful because it allowed the team to closely study and understand user needs before investing in technological innovations to automate the process.

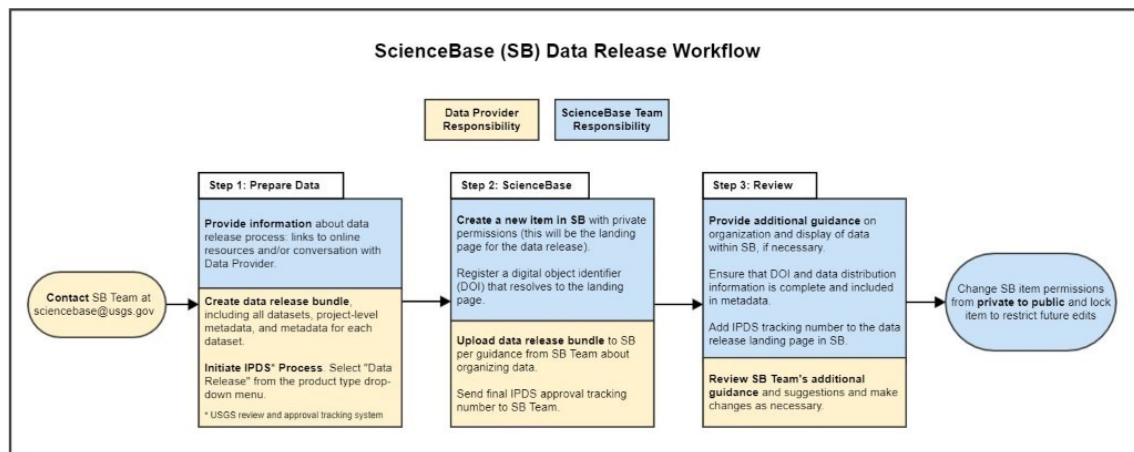


Figure 2. ScienceBase data release workflow diagram version 1, used during the first year of ScienceBase operation. Each step required manual processing and review by both the data publication author and the SDM team.

In the first step, a data author contacted the SDM team to start the process. The team then shared information with the author about workflow steps and recommendations specific to their data release. Next, the team manually created a new landing page in ScienceBase to host data files and reserved a unique alphanumeric Digital Object Identifier (DOI) to use as a persistent identifier in the data release's citation and metadata. DOIs were reserved through the USGS DOI Tool, an internal application that interfaces with DataCite⁵ to assign unique digital identifiers.

In the next step, the author uploaded data and metadata files to their landing page in ScienceBase. USGS data release policy requires that all data be accompanied by one or more metadata files in XML (Extensible Markup Language) format. The metadata files must adhere to the structure of an established standard, either the Federal Geographic Data Committee (FGDC) Content Standard for Geospatial Metadata (CSDGM) or International Organization for Standardization (ISO) standard (USGS, 2017b). At the time the SDM team developed the data release workflow, ScienceBase already had the capability to parse metadata in these two formats: the ScienceBase system can map key elements from uploaded metadata files and automatically populate fields on a landing page. The original metadata is also stored as a file attachment on the landing page. Authors and the SDM team can provide additional information or labels on the landing page as needed.

The SDM team focused on the standardization of data release content fields early in the workflow development process. When standardized, fields such as recommended citation, publication date, persistent identifiers, and keywords can enable targeted queries, tracking, reporting, specialized display, and efficient management of data releases as a collection in the repository. Standardizing content allowed the SDM team to support repository business query and reporting needs (e.g., reporting and displaying products by science center, and finding products by unique identifier and author ID).

Lastly, the SDM team ran a set of manual checks against a quality control checklist before making data releases public. The quality control checklist included formatting standards (e.g., placement and format of the data release citation), organizational best practices (e.g., a set of options for structuring data releases in a clear and accessible way),

5 DataCite: <https://datacite.org>

and compliance with USGS policy (e.g., verification of review and approval for release of the data).

Assessment Tools

In 2016 and 2017, the SDM team performed three separate assessments to help inform the development of the data release workflow, resulting in an iterative improvement to the application.

The SDM team participated in a Data Management Maturity Assessment of ScienceBase, conducted by the American Geophysical Union (AGU) in 2016 (Stall et al., 2016). The week-long assessment provided the team with a detailed review of the existing ScienceBase system and data release operations. The resulting report identified several areas for the SDM team to focus on as high priority goals. Recommendations included documenting the data management strategy for data release, developing a communication plan to interact with stakeholders, and continuing to improve the ScienceBase technology stack and processes that support file handling and preservation.

Additionally, during the first phase of workflow development, the SDM team regularly collected input and recommendations from new users via a short online feedback form. This feedback helped the team track user challenges and successes. For example, one challenge was that information about the new data management policies had initially been shared inconsistently with researchers, due to the distributed nature of the USGS as an organization. Positive feedback about the SDM team's responsiveness and ability to assist users was also noted, reinforcing the team's commitment to maintain reliable user support, even as automated steps were added to the data release workflow.

Lastly, using best practices for usability testing (digital.gov team, n.d.), the SDM team conducted a formal usability study on the ScienceBase system in 2017. The team identified specific data release tasks for participants, USGS scientists new to ScienceBase, to complete. They recorded various routes that users took to try to accomplish these tasks, focusing on areas where participants expressed confusion or encountered roadblocks. Conclusions from this assessment resulted in recommendations for updates to the ScienceBase user interface to clarify and further streamline the process for users.

Planning Next Steps

The three assessments conducted by the SDM team helped the team plan and prioritize future efforts. The identified goals were:

- meet the requirements of the USGS Trusted Digital Repository certification;
- improve workflow usability for authors;
- ensure that SDM team members provide a consistent level of user support amid growing demand;
- continue to refine and expand the ScienceBase data release process in a scalable, sustainable manner for use across the Bureau.

These goals were accomplished through both non-technical improvements (e.g., training and outreach) and technical improvements (e.g., system integration with other data management applications, new features, and automation of key steps in the

process). They were also informed by several Earth science data communities, including Earth Science Information Partners⁶ and the Research Data Alliance⁷. SDM team members regularly engaged with these communities, gaining up-to-date information about data publication practices, current technologies, and developing guidance for data citation. The improvements, described next in this paper, resulted in a clearer and more sustainable process for USGS scientists to formally release data.

Outreach and Communication

Following the recommendations of the AGU Data Management Maturity Assessment (Stall et al., 2016), the SDM team wrote a set of documents to support internal data release processes. The documents included a data management strategy, standard operating procedures, a ScienceBase user agreement, and a communication and outreach strategy. These documents were helpful in defining the scope of the team's work and helped the team plan for a projected increase in the rate of data release publication.

In a distributed organization, such as the USGS, it is important to consider measures such as broad communication and community buy-in for new initiatives. To address this, the SDM team's communication plan outlined a series of steps to help the team improve information dissemination and customer service across the Bureau.

In 2016, the SDM team partnered with the USGS Office of Science Quality and Integrity (OSQI) to organize approximately 25 in-person and remote presentations to more than 1,200 scientists across the Bureau. The presentations increased awareness of the new data release policies and available implementation methods, including the ScienceBase data release process. Additional methods of outreach to the Bureau included newsletter-style update emails, direct assistance for authors and data managers, webinars, and guidance published to the USGS data management website⁸.

The SDM team also initiated the ScienceBase Data Release Network (the Network), a group of data managers from science centers across the USGS. The SDM team reached out to USGS science centers and programs to invite appropriate staff to participate. By 2019, the Network participant list included over 100 USGS personnel. An important goal for establishing the Network was to engage directly with data managers, who often assist USGS scientists with data release processes. In this role, they disseminate information about the ScienceBase workflow and USGS policy requirements to their respective science centers and programs. The Network also proved to be an excellent source of feedback for the SDM team, fostering connections with data managers who are highly involved in the data release process and can share valuable perspectives.

The SDM team developed guidance materials specifically for the Network, to enable participants to support data authors in the same manner as the small SDM team. This helped distribute the effort in assisting scientists across a distributed organization and prevented potentially time-sensitive bottlenecks in data release. Information provided by the SDM team to the Network included a webinar training for new participants (repeated on a monthly basis), an online wiki with a shared collection of helpful resources including demo videos, quality control checklists, workflow diagrams, and answers to FAQs. The SDM team also held monthly webinars as a forum for Network participants to ask questions, provide feedback, share updates, and engage in discussions.

⁶ Earth Science Information Partners (ESIP): <https://www.esipfed.org>

⁷ Research Data Alliance (RDA): <https://www.rd-alliance.org/>

⁸ USGS Data Management: <https://www.usgs.gov/products/data-and-tools/data-management>

These conference calls helped create a sense of community and a shared understanding of the ScienceBase data release process.

Lastly, the SDM team responded to requests from the Network to consolidate disparate sources of information about the data release process and developed a focused guidance webpage⁹ with step-by-step instructions and FAQs. The team regularly updated this page as needed based on input collected from the online feedback form and the Network.

These steps in documentation, communication, and creating community were essential for maintaining the SDM team's efficiency and responsiveness to user needs. The steps helped the team reduce the number of one-on-one meetings with data authors and enabled the authors to more independently work through the data release process.

Technical Improvements: A Phased Approach to Automation

As usage of the ScienceBase repository increased, the SDM team focused on creating efficiencies for both users and the SDM team. In the first iteration of the data release workflow, shown earlier (Figure 2), the supporting processes were completed manually by the SDM team. The manual workflow was used for approximately one year, a period during which the team identified tasks in the workflow that could be automated. To move towards a more streamlined process, a four-phase plan was initiated to automate many of the manual steps in the workflow.

Phase 1: Leveraging Google™ Drive¹⁰ applications

The SDM team created an online entry form in Google Forms to collect information from data authors. The form collected all the information needed to create a new data release landing page in ScienceBase. The information was automatically stored in a Google Sheets spreadsheet, and an automated notification email was sent to the SDM team to inform them about a new submission. Using the information submitted through the form, the team manually created a landing page in ScienceBase and reserved a DOI for the authors.

This workflow update improved efficiency for the SDM team because all information was collected and stored automatically in Google Drive and one-on-one conversations with data authors were no longer required. In version 2 of the workflow diagram (Figure 3), step 1 was effectively streamlined and no longer involved manual work by the SDM team.

⁹ ScienceBase Data Release Instructions: <https://www.sciencebase.gov/about/content/data-release>

¹⁰ Google Drive: <https://www.google.com/drive/>

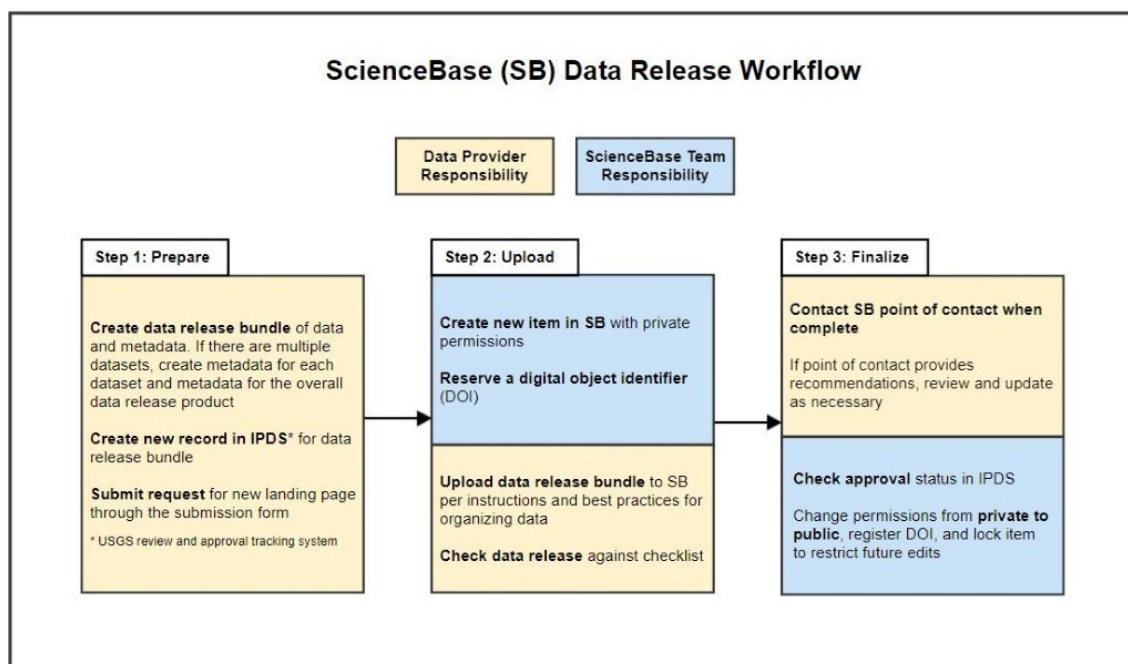


Figure 3. ScienceBase data release workflow diagram version 2. The addition of a submission form to collect information from authors resulted in fewer steps performed by the SDM team (compared to workflow version 1).

Phase 2: Application development

Although the Google Drive application functioned well, the SDM team decided to create additional efficiencies and functionality by developing a more customized application, the ScienceBase Data Release (SBDR) Tool¹¹. The Tool fully automates the creation of new landing pages and DOI reservation. This level of automation is possible due to the robustness of the ScienceBase and USGS DOI Tool APIs; all actions that users can perform through a web browser, such as creating and editing landing pages, can also be accomplished through code.

In the SBDR Tool, a data author fills out a short form to describe their data release (Figure 4). The Tool then connects to ScienceBase and the DOI Tool, via their respective APIs, to automatically create a new landing page and reserve a DOI. The landing page stores all the information the SDM team needs to finalize the data release at the end of the process. In addition, the new landing page URL is passed back to the DOI Tool as the location to which the DOI will resolve once it is published. After submitting the form through the SBDR Tool, the data author receives a notification email containing a link to the new landing page, the reserved DOI, and instructions about next steps.

¹¹ ScienceBase Data Release Tool: <https://www.sciencebase.gov/datarelease>

Create a New ScienceBase Data Release

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.

*** 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.

*** Would you like ScienceBase to assign a digital object identifier (DOI) for your data release?**

Yes
 No, I already have one

Data Release Authors
If yes, please add the authors of this data release. ORCID field is optional.

First Name	Last Name	Middle Initial	ORCID	
<input style="width: 90%;" type="text" value="Jane"/>	<input style="width: 90%;" type="text" value="Doe"/>	<input style="width: 90%;" type="text" value=""/>	<input style="width: 90%;" type="text" value="0000-0000-0000-0000"/>	Remove Data Author

Figure 4. Input form of the ScienceBase Data Release Tool (note: some fields refer to internal USGS systems and terminology; the Open Researcher and Contributor ID (ORCID) is an external persistent digital identifier required for USGS authors).

Phase 3: Python scripts to streamline quality assurance/quality control and publication

The SBDR Tool streamlined steps 1 and 2 in the data release workflow (Figure 5). In order to streamline step 3 of the workflow, the SDM team developed a set of Python scripts maintained as a Jupyter Notebook (Kluyver et al., 2016). These scripts connect to the ScienceBase API and programmatically run through a set of quality control checks, finalize the data release, and make the data release publicly available. The steps that

were upgraded from manual to automated include the following: validating metadata, checking identifier format and location, standardizing citations across all pages of a data release, updating permissions to make the data release public, adding a data release label, and generating emails to authors and data managers to inform them when data releases have been made public. As a result, the checks and finalization steps that previously were performed manually by the SDM team are now accomplished programmatically (with the exception of a few minor visual checks that were not possible to automate) (Figure 5, step 3).

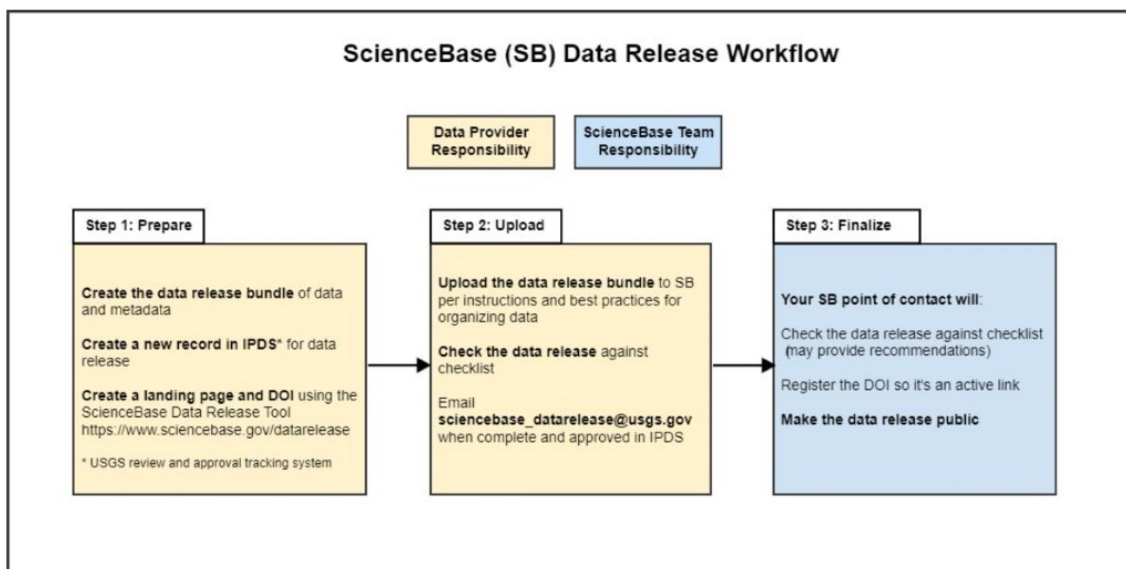


Figure 5. ScienceBase data release workflow diagram version 3. As a result of additional automation, the SDM team no longer needs to manually create landing pages and DOIs for data authors (step 2), and almost all checks and finalization steps are completed programmatically using Jupyter Notebook scripts (step 3). The SDM team is only involved in step 3 of the process.

Phase 4: Customization options to facilitate data manager support

In the fourth phase of technical improvements, the SDM team addressed the needs of another type of user involved in the data management process – the data manager. USGS policy (USGS, 2017a) requires science centers and programs to have a dedicated data manager to assist scientists with the data release process. The SBDR Tool and Python finalization script were designed with customizable features that can help data managers perform their tasks more effectively. The SDM team reached out to USGS data managers, via the ScienceBase Data Release Network, to discuss and select specific customization options for science centers.

Custom actions in the SBDR tool are triggered when a data author selects a specific science center in the tool’s input form. Among possible actions, the tool can:

- automatically notify a data manager when an author from their center has started a new data release;
- grant read/write permissions to data managers for new landing pages in ScienceBase;

- grant read/write permissions for the assigned DOI in the USGS DOI Tool;
- create landing pages in a specified location within ScienceBase, allowing data managers to more easily track and manage data releases from their center.

For the final step of the process, data managers can request that data releases from their center be moved to a specific location in ScienceBase for publication. This can further enable management and tracking of the releases. Additionally, data managers have the option to receive a notification email when a data release from their center has been published.

Following implementation of the four development phases, the data release process became more efficient for data authors, data managers, and the SDM team. Data authors receive new landing pages and DOIs immediately upon submission of the SBDR Tool's online form, and no longer need to wait for a response from the SDM team. Data managers can customize key steps in the process to more easily manage collections of data releases. The SDM team no longer creates landing pages manually and can complete finalization steps faster and with more accuracy. As a result, the SDM team has been able to successfully increase the rate of data release publication without also increasing the work hours spent supporting the process (Figure 7, discussed later in the paper). The time needed to complete a data release can vary widely due to the multi-step process, which can involve communication exchanges between data authors and the SDM staff. Because of this, quantitative measures of efficiency are difficult to produce accurately. In general, a data release performed manually could take up to several days to complete; however, with automated processes, the release can be completed within an hour.

Additional technical improvements

As the SDM Team developed and updated the ScienceBase data release workflow, they identified additional technical improvements to improve efficiency. These included system integration with other USGS tools, tracking and metrics, and improved access for data reviewers.

System integration

The SDM team manages or interacts with multiple tools required for USGS data release. Increasing the technical integration of these systems has been central to the SDM team's development plan for an efficient data release workflow.

In addition to ScienceBase and the USGS DOI Tool, the team also manages the USGS Science Data Catalog¹², a metadata catalog referencing all public USGS data products. According to USGS policy (USGS, 2017b), metadata records describing all published data products must be deposited in and shared through the Science Data Catalog (SDC). Submitting metadata records to the SDC was initially a manual step for data authors and managers, but the SDM team added the ability to automatically send metadata records to the SDC directly from published data releases in ScienceBase. This was accomplished by creating an Open Archives Initiative – Protocol for Metadata Harvesting (Lagoze et al., 2015) connection to harvest and transfer metadata between ScienceBase and the SDC. Data authors can now meet the policy requirement without completing additional manual steps when they use ScienceBase for data release.

Another example of system integration was the development of a connection between ScienceBase and the USGS DOI Tool. The data release workflow contains a

¹² USGS Science Data Catalog: <https://data.usgs.gov/datacatalog>

step in which the SDM team updates the status of data release DOIs from ‘reserved’ to ‘published’. This publishes the DOIs to DataCite and ensures that DOI URLs are active links to their associated landing pages. The team initially completed this step manually by using the USGS DOI Tool. A feature was added to ScienceBase to allow the SDM team to connect to the DOI Tool via its API, and as a result, the team can now publish DOIs directly from ScienceBase.

Lastly, the SBDR Tool was updated to allow connection with a web service endpoint of the Information Product Data System (IPDS), the internal product review and approval tracking system of the USGS. All USGS products, including data releases, are required to have an associated entry created in IPDS (USGS, 2016b), which stores much of the information that will later be entered in the SBDR Tool. The SBDR Tool offers users the option to pull content from the associated entry in IPDS by using an autofill feature, saving time for the authors and standardizing information across systems.

Queries, tracking, and metrics

As the number of data releases in ScienceBase increased, the ability to locate and track data releases became an important goal in order to help with the increasing amount of analytics data being captured by the system. A new type of administrative tag was added in ScienceBase to identify formal USGS data release products. The tag is programmatically added to all data releases as they are finalized, providing the ability to query and quantify all public data products in ScienceBase.

USGS authors are required to obtain an Open Researcher and Contributor ID (ORCID) before publishing scientific papers or releasing data (USGS, 2016b). The intent of this requirement is to improve the ability to track an author’s scientific output and provide credit if data are reused. ScienceBase stores ORCIDs in a standardized field to enable search queries and to improve discoverability of data products produced by individual scientists.

ScienceBase landing pages were updated to include a metrics display (Figure 6). This provides data authors with a way to track quantitative metrics by date ranges. For example, authors can see the number of people who viewed their data releases and downloaded the associated files (in the File Downloads section).



File Downloads

Filename	Total Downloads
ZIP (all files)	81

Figure 6. An example of the metrics display for a ScienceBase data release.

Access options and supporting data review

When USGS authors publish in external journals, the data supporting the publications may need to be made available for a journal's selected reviewers prior to publication. During this stage of the publication process, however, the data are not always publicly available yet. To facilitate access for non-USGS reviewers, a new feature was added to ScienceBase: data authors can generate temporary access URLs to share with reviewers. The access URLs allow reviewers to view permission-controlled data without the need to authenticate into the ScienceBase system.

File transfer, storage, and backup considerations

Technical limitations on transferring and storing large data files is an ongoing challenge in the scientific community. To help address this challenge, the SDM team increased the file size limit for uploads and downloads through the ScienceBase user interface from 2GB to 10GB. Additionally, a new data transfer option was introduced to better accommodate large files (e.g., over 200GB), which are increasingly more common for data release products the SDM team supports. Although large file transfer remains a challenge due to USGS network bandwidth limitations and security considerations, the SDM team plans to continue to expand file transfer support with more advanced file transfer protocols and modern cloud-based capacities for virtually unlimited scaled growth.

Usability improvements

One of the most significant improvements in usability was the creation of the ScienceBase Data Release (SBDR) Tool, described earlier in this paper. Because ScienceBase is a complex application that can present usability challenges for authors creating landing pages on their own, the development of a tool that automates the process proved to be very helpful in reducing confusion and training needs. Additionally, the ScienceBase user interface was updated to make it more intuitive, based on recommendation from the usability test. The updates included improving clarity of the text and moving key user actions so that they are more prominently displayed on landing pages.

As the usage of the ScienceBase data release process increased, system integration, usability improvements, and accommodations for large data files were important ways in which efficiencies were achieved to improve the user experience.

Outcomes and Future Directions

ScienceBase Achieved Trusted Digital Repository Status in USGS

The certification process for USGS Trusted Digital Repositories (TDRs) requires a repository team to submit a formal application based on the Core Trust Seal criteria (Edmunds et al., 2016). The submission is then reviewed by the USGS Fundamental Science Practices Advisory Council Subcommittee on Data Preservation, which evaluates and establishes TDRs for the Bureau. ScienceBase was designated a USGS TDR in early 2017 and was the first USGS data platform to achieve this status. This achievement was an important milestone for the SDM team and ensured that USGS scientists have an approved location for publishing their data products.

Results of Streamlining and Scalability Efforts

Since it was initiated in 2015, the ScienceBase data release process has seen wide adoption across the Bureau. By the end of fiscal year 2020, a total of 97 USGS science centers and programs (out of 116) had completed at least one data release in ScienceBase, and a total of 3,805 data release products had been published in ScienceBase.

The annual (Figure 7) and average monthly rate of USGS data release publication has steadily increased. In FY 2016, the SDM team published an average of 14 data releases per month; in FY 2020 that number had increased to 95 data releases per month.

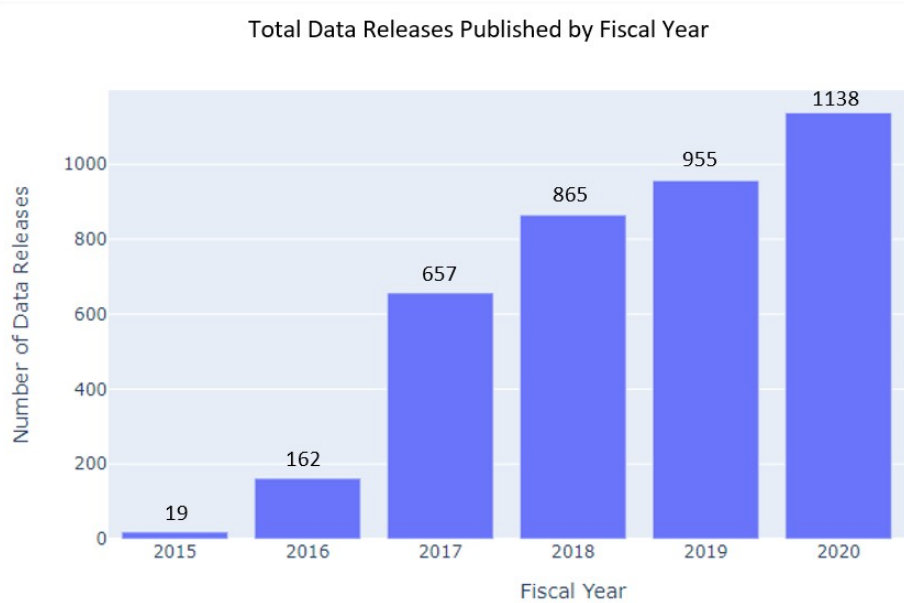


Figure 7. Number of new USGS data releases published in ScienceBase by fiscal year.

This increase was accomplished without corresponding growth in the size of the SDM team, which continues to allocate approximately two full-time staff to assist with data release. The team has been able to maintain a quick turnaround time for all data release requests and questions; while the SDM team officially offers a two-day turnaround time, the majority of releases are completed within one day. The team attributes this scaling success to the carefully planned, iterative steps that have automated and streamlined the data release process.

Impact of Scientific Data Released in ScienceBase: Example

The data product ‘Map data of landslides triggered by the 25 April 2015 Mw 7.8 Gorkha, Nepal earthquake’ was published in ScienceBase in 2017 (Roback et al., 2017). This data release, created in cooperation with USGS partners at the University of Michigan, the Swiss Federal Institute of Technology, and in Nepal, mapped earthquake-triggered landslides using high-resolution (<1m pixel resolution) pre- and post-event satellite imagery. Since 2017, the data have been cited by five publications, including one that used the data in a study proposing a method for near real-time prediction of earthquake-induced landslides (Tanyas et al., 2019). This data release is an example of how use of the ScienceBase repository and its data release workflow can facilitate data reuse, data citation, and open science best practices.

Future Directions

The SDM team continues to develop updates to the ScienceBase repository to improve user experience, address changing data needs in the Bureau, and increase efficiencies for the small repository management team.

Improved search and browse for data release products

The SDM team is pursuing an improved user interface (UI), including a more attractive and intuitive access point for data products, to better enable search and browse capabilities. Using standardized descriptive information, consistent structure, and controlled use of labels, it is possible to work with the data release products as a collection and implement rules for how they can be filtered, explored, and displayed. The SDM team will also build on existing efforts to provide analytics and visualization options to accompany user experience with data release products.

Tracking downloads and downstream use of USGS data

The SDM team will continue to refine a process for tracking use of ScienceBase data releases, as they are cited in formal publications and other outlets such as scientific software. In 2019, SDM team members explored a workflow to track occurrences of USGS data products, writing Python scripts to discover where the unique USGS DOI prefix (10.5066) appears in peer-reviewed literature. This will become an automated process that regularly updates USGS data release products with information about the associated publications where they are cited. Capturing and storing this information in ScienceBase adds value for data authors and the public and helps demonstrate the value of data citation and best practices for data management.

Expanded file support and value-added features

As new cloud-based technologies become standard components of the technical architecture in modern data platforms, new services and features are being developed in ScienceBase that can offer solutions for some of the traditional challenges in data management. These include improved options for transferring large files, and more sophisticated methods of access that, for example, allow users to request a subset of a large file rather than downloading the entire dataset. The SDM team continues to create more efficient and practical options for users to store and serve files in ways that maximize utility.

Guiding and implementing best practices for information systems in the USGS

In the process of transitioning an existing application into a USGS Trusted Digital Repository and honing an efficient data release process, the SDM team learned the importance of encouraging the broader USGS organization to follow technical best practices in information management. These best practices can include maintaining integrity of API endpoints, supporting linkages via primary keys, and using technical design techniques such as standardized labels to support parameterized search across systems. Coordinating best practices across enterprise systems can enable nimble system response when using code to share content.

The SDM team will continue to help define and implement robust conventions to improve how information is structured in and retrieved from primary information systems in the USGS. The team will also identify opportunities to promote these best

practices in the USGS, thus continuing to contribute to Bureau efforts to improve programmatic connection of enterprise information systems.

Lessons Learned

In reviewing the work completed over five years as described in this paper, the SDM team identified several key actions that helped it efficiently support data publication across a distributed Federal science agency:

- **leveraging an existing application**, instead of creating a new data repository, to save costs in acquisition and migration of resources, and to retain a system already familiar to users;
- **strategic planning**, prior to the start of workflow design and technical development, to incorporate key goals and user input early in the process, decreasing the number of necessary iterations in development;
- **engaging** with data communities of practice to stay up to date on standards, best practices, and emerging trends regarding data publication and repositories, helping users experience the most modern data management practices in repositories;
- **envisioning a clear, easy workflow** for data authors to follow, reducing barrier to entry and boosting user confidence in the repository and publication process;
- **performing tasks manually** during initial stages of workflow implementation to help identify potential steps to automate;
- **including scalability considerations** early in the planning process to prepare for future growth;
- **writing internal guidance documents** (e.g., data management strategy, communication strategy, standard operating procedures) to standardize and define the scope of user support processes, improving alignment and consistency;
- **using iterative design** to allow flexibility in addressing needs as they arise;
- **anticipating query and metrics needs** to ensure that content necessary for reporting is collected during the process;
- **standardizing information** in metadata to enable programmatic queries and edits, allowing for development of automated processes
- **automating data curation tasks** to reduce the time spent supporting the process and improve quality control and standardization;
- **connecting relevant, separate applications via APIs** to streamline the process for both data authors and the repository team;
- **collecting user feedback** to identify areas for improvement and to better understand user needs;

- **testing usability** and applying lessons learned to the workflow and technical design;
- **employing multiple communication methods** to reach dispersed user groups, because users learn and obtain information through various mechanisms;
- **establishing a community of users** (e.g., the ScienceBase Data Release Network) and maintaining regular communication to receive up-to-date user feedback to improve usefulness;
- **applying for certification** as a Trusted Digital Repository to increase confidence and reputation across the organization.

Additional factors that contributed to operational success:

- **coordinated management of applications:** tool integration and shared development goals can be more efficiently aligned if the data management tools involved in the process are under one management umbrella (in this case the USGS SAS Science Data Management team);
- **data management policies within the organization:** USGS policies that formally document requirements for scientists provide a critical foundation for this effort by defining a scope and clear process around which to design a technical workflow. Policies also convey that new requirements are backed by a high level of authority within the USGS, allowing the SDM team to interact with authors as facilitators in the data release process rather than enforcers or auditors.

Conclusion

The SDM team applied for and received the designation of USGS Trusted Digital Repository for ScienceBase in 2017, achieving its goal to transition the ScienceBase platform to its current status as an official USGS repository. The USGS is a distributed organization, reliant on scientists and data managers across geographically separated science centers to actively perform data management for the scientific data they generate. The SDM team set out to provide the Bureau with a Trusted Digital Repository that supports official release of USGS data products across the science centers. This ambitious goal was accomplished by creating a manual process that later translated into a set of automated tasks, increasing efficiencies for both the repository's users and the small team that manages the repository. Also important in the approach was creating a solid community around the system and drawing upon the feedback from that community to successfully guide critical advancements in the data release workflows and technical functionality of ScienceBase. Additionally, incorporating lessons learned from multiple assessments, both internal and external, contributed to the SDM team's ability to transform ScienceBase into a focused and useful application for the Bureau. The success of this transition is measured by the steady increase in usage of ScienceBase by scientists across the USGS; by the end of fiscal year 2020, the number of official data releases had exceeded 3,500 and nearly every USGS science center had published data using the recommended workflow. By describing the USGS experience in leveraging an

existing application to meet broad Federal data directives, we hope to offer a helpful roadmap for other organizations considering similar transitions.

Acknowledgements

We would like to thank the following for their contributions to our paper: Kelly Haberstroh, USGS; Emily Read, USGS; and Amy Forrester, University of Tennessee. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

- Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., Uhler, P., & Wouters, P. (2004). An international framework to promote access to data. *Science*, 303(5665), 1777–1778. doi:10.1126/science.1095958
- Burwell, S. M., Mancini, D. J., Park, T., & VanRoekel, S. (May 9, 2013). *Open Data Policy —Managing Information as an Asset* (Memorandum M-13-13). Executive Office of the President, Office of Management and Budget. Retrieved from <https://obamawhitehouse.archives.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf>
- Defne, Z., & Ganju, N. K. (2018). *Shoreline change rates in salt marsh units in Edwin B. Forsythe National Wildlife Refuge, New Jersey* [Data set]. U.S. Geological Survey data release. doi:10.5066/F7PN94K2
- Digital.gov team, U.S. General Services Administration (GSA) Technology Transformation Service (n.d.). *Planning a Usability Test*. Retrieved from <https://www.usability.gov/how-to-and-tools/methods/planning-usability-testing.html>
- Edmunds, R., L'Hours, H., Rickards, L., Trilsbeek, P., Vardigan, M., & Mokrane, M. (2016). *Core Trustworthy Data Repositories Requirements*. Zenodo. doi:10.5281/ZENODO.168411
- Exec. Order No. 13642. (May 9, 2013). *Making Open and Machine Readable the New Default for Government Information*. 78 Fed. Reg. 28111. <https://www.federalregister.gov/d/2013-11533>
- Fundamental Science Practices Advisory Committee. (2011). *U.S. Geological Survey Fundamental Science Practices*. U.S. Geological Survey Circular 1367. doi:10.3133/cir1367

- Holdren, J. P. (2013, February 22). *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. Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf
- Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, 29(4), 258–268. doi:10.1080/10580530.2012.716740
- Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J. B., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C., & Jupyter Development Team. (2016). Jupyter Notebooks – a publishing format for reproducible computational workflows. In F. Loizides & B. Schmidt (Eds.), *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (pp. 87–90). IOS Press. doi:10.3233/978-1-61499-649-1-87
- Kriesberg, A., Huller, K., Punzalan, R., & Parr, C. (2017). An analysis of federal policy on public access to scientific research data. *Data Science Journal*, 16(0), 27. doi:10.5334/dsj-2017-027
- Lagoze, C., Van de Sompel, H., Nelson, M. & Warner, S. (Eds.). *Open Archives Initiative Protocol for Metadata Harvesting* Version 2.0 of 2002-06-14. (2015). Retrieved from <http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm>
- Latysh, N., Kirk, K. G., & Faundeen, J. (2020). *Purpose and benefits of U.S. Geological Survey Trusted Digital Repositories*. U.S. Geological Survey Fact Sheet 2020–3032. doi:10.3133/fs20203032
- Molloy, J. C. (2011). The Open Knowledge Foundation: Open data means better science. *PLoS Biology*, 9(12), e1001195. doi:10.1371/journal.pbio.1001195
- Roback, K., Clark, M. K., West, A. J., Zekkos, D., Li, G., Gallen, S. F., Champlain, D., & Godt, J. W. (2017). *Map data of landslides triggered by the 25 April 2015 Mw 7.8 Gorkha, Nepal earthquake* [Data set]. U.S. Geological Survey data release. doi:10.5066/F7DZ06F9
- Sheehan, J. (2016, February 22). *Increasing Access to the Results of Federally Funded Science*. The White House Blog. Retrieved August 8, 2018 from <https://obamawhitehouse.archives.gov/blog/2016/02/22/increasing-access-results-federally-funded-science>
- Stall, S., Hanson, B., & Wyborn, L. A. I. (2016). The American Geophysical Union Data Management Maturity Program [Abstract]. *Abstracts with Programs*, 48(7). Geological Society of America. doi:10.1130/abs/2016AM-284514
- Tanyas, H., Rossi, M., Alvioli, M., van Westen, C. J., & Marchesini, I. (2019). A global slope unit-based method for the near real-time prediction of earthquake-induced landslides. *Geomorphology*, 327, 126–146. doi:10.1016/j.geomorph.2018.10.022

- U.S. Geological Survey. (2016a). *Public Access to Results of Federally Funded Research at the U.S. Geological Survey*. Office of Science Quality and Integrity. Retrieved August 8, 2018 from <https://www.usgs.gov/about/organization/science-support/science-quality-and-integrity/public-access-results>
- U.S. Geological Survey. (2016b). *Survey Manual 502.4 - Fundamental Science Practices: Review, Approval, and Release of Information Products*. Retrieved August 1, 2018 from <https://www.usgs.gov/about/organization/science-support/survey-manual/5024-fundamental-science-practices-review-approval>
- U.S. Geological Survey. (2017a). *Survey Manual 502.6 – Fundamental Science Practices: Scientific Data Management*. Retrieved August 1, 2018 from <https://www.usgs.gov/about/organization/science-support/survey-manual/5026-fundamental-science-practices-scientific-data>
- U.S. Geological Survey. (2017b). *Survey Manual 502.7 - Fundamental Science Practices: Metadata for USGS Scientific Information Products Including Data*. Retrieved August 1, 2018 from <https://www.usgs.gov/about/organization/science-support/survey-manual/5027-fundamental-science-practices-metadata-usgs>
- U.S. Geological Survey. (2017c). *Survey Manual 502.8 - Fundamental Science Practices: Review and Approval of Scientific Data for Release*. Retrieved August 1, 2018 from <https://www.usgs.gov/about/organization/science-support/survey-manual/5028-fundamental-science-practices-review-and>
- U.S. Geological Survey. (2017d). *Survey Manual 502.9 - Fundamental Science Practices: Preservation Requirements for Digital Scientific Data*. Retrieved August 1, 2018 from <https://www.usgs.gov/about/organization/science-support/survey-manual/5029-fundamental-science-practices-preservation>
- Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1). doi:10.1038/sdata.2016.18