# Data Release Use Case Team: Recommendations

April 2, 2013

## Introduction

In March 2012, the Data Release Use Case Team convened through the USGS Community for Data Integration (CDI) Data Management Working Group to develop a procedure that enables USGS employees to determine if a particular set of data is approved for release.

The team is comprised of representatives from various mission areas, including Climate and Land Use Change, Core Science Systems, Energy and Minerals and Environmental Health, and Natural Hazards, and the Office of Science Quality and Integrity. The skill set of the nine-member team includes data experts, approving officials, software engineers, and experts in information modeling and USGS policy (refer to Appendix A for the members list). The team used a semantic development (Use Case) approach, facilitated by Dr. Peter Fox of Rensselaer Polytechnic Institute (RPI). More information on the team's activities is available on the CDI Confluence site at https://my.usgs.gov/confluence/display/cdi/Data+Release+Use+Case+Team.

The team's work involved developing use case diagrams that depict the current and or recommended flow of the review and approval process for various data release scenarios. The first (referred as "primary") use case outlines the workflow for release of data in a USGS publication series (typically the USGS Data Series or Open-File Report series) and is based on current USGS policy (refer to SM 1100.3). Supporting documentation that clarifies USGS procedures for data release is included.

The development of use cases covering other USGS data release scenarios, such as review and approval processes for data released on the Web (and not published in a USGS series publication) is underway by the team.

This document provides recommendations for improving the current review and approval processes. The recommendations address challenges the team identified in determining whether USGS legacy data and data associated with an approved publication or released in a USGS series can be disseminated. If implemented, the recommendations will improve the processes and procedures for public dissemination of USGS data.

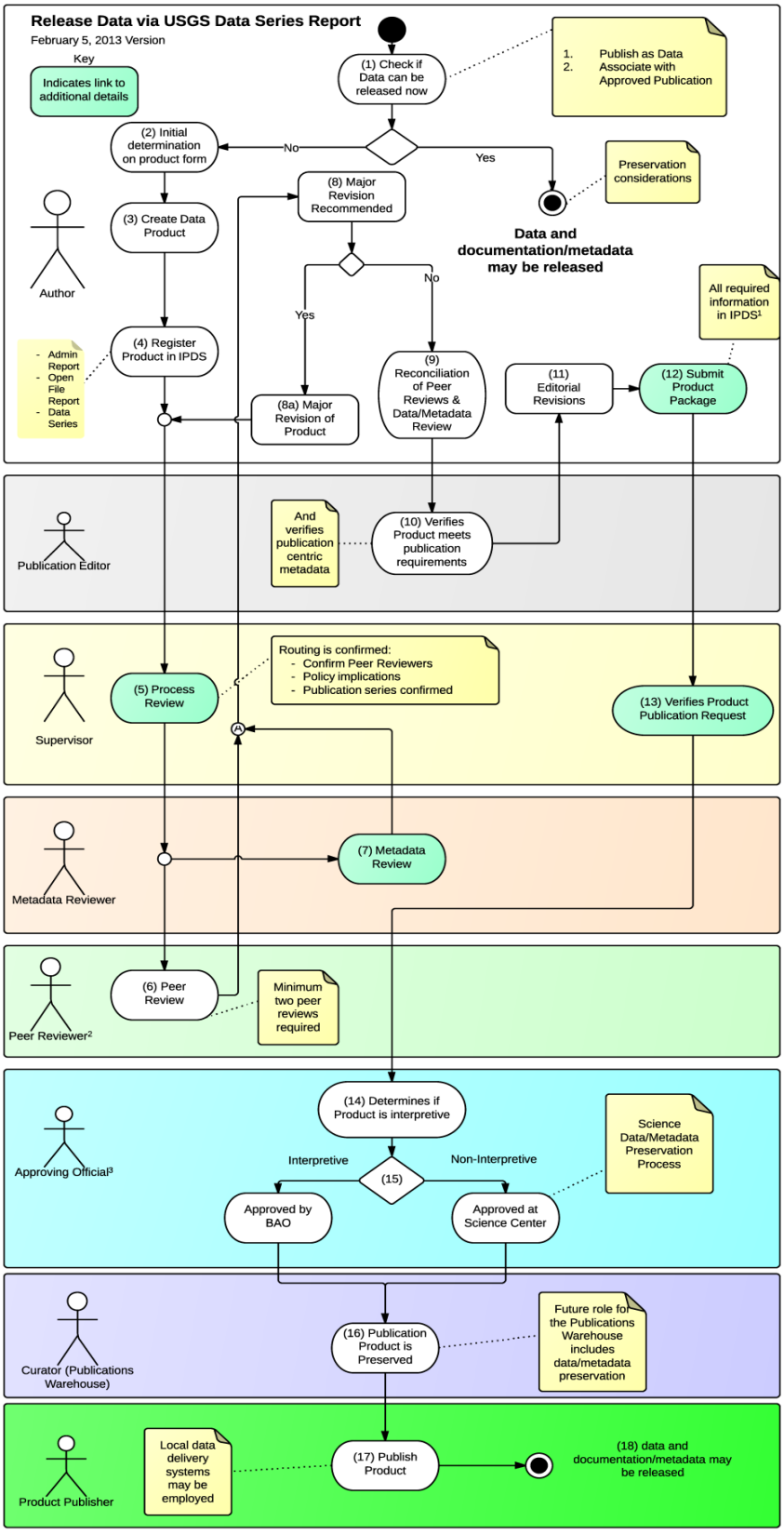


**Figure 1.** Primary Use Case Diagram portraying the responsibilities of individuals or entities within the USGS for releasing data via a USGS series publication.

## Challenges and Recommendations

The use case depicts the scenario of a USGS scientist whose goal is to release data through a USGS series publication.

Figure 1 shows the steps of the workflow. Detailed descriptions of the steps are in Appendix B. The challenges are listed in the order they were encountered in the workflow and the associated steps are referenced with each challenge (Figure 1). In developing the use case the team realized there are multiple challenges that impede the scientist's ability to easily complete the process of data publication. The following recommendations are in response to these challenges, with intent to improve the process of releasing USGS data.

**Challenge:** The author needs to check if the data have previously been published or are associated with an approved publication. Currently, there is no comprehensive system in USGS to link data collections to publications, therefore the author needs to search individual publication catalogs and peruse publications to determine if the data have already been published or are associated with an approved publication. This process is time-consuming, error-prone, and not compatible with automation. (Workflow Step 1)

**Recommendation # 1.** Implement a USGS system for assigning Digital Object Identifiers (DOIs) to both digital data and publications as part of the approval process.

**Recommendation # 2.** Develop search capabilities for USGS publication catalogs that include DOIs and the ability to link associated products between catalogs.

**Challenge:** The author decides to publish the data in a Data Series or Open-File Report, and proceeds to develop the product. This step is challenging because some publication formats (such as PDFs) are incompatible with the expectation that the digital data be made available for data reuse and integration. Therefore, USGS scientists are experimenting with new publication formats such as online reports that use html and xml for releasing digital data. Unfortunately, these new publication formats are not handled particularly well by existing USGS publication resources, especially when large volumes of data are published. (Workflow Steps 2 and 3)

**Recommendation # 3.** Implement USGS online data services that allow approved data to be released in more useful formats. Where appropriate, maintain the connection between the data and the approved publication(s); this can be facilitated by effective use of DOIs.

**Recommendation # 4.** Provide training and guidance on options and standards for useful data formats and how to prepare data in these formats for release on the Web.

**Challenge:** When the author registers the product in the USGS Information Product Data System (IPDS), peer reviewers are designated and verified by the supervisor. Peer reviewers then review the product

and the underlying data for accuracy and usefulness. However, the USGS policy is unclear about the responsibilities of peer reviewers to review the data for data products. (Workflow Steps 4, 5, and 6)

**Recommendation # 5.** Develop guidance to aid selection of appropriate peer reviewers who have qualifications to review a specific type of data (e.g., a GIS expert should be reviewing a GIS data).

**Recommendation # 6.** Develop guidance on the responsibilities of 1) peer reviewers to review data products, 2) peer reviewers of interpretive publications to review the data underlying interpretive products, and 3) authors to provide sufficient information to enable peer reviewers to meet their responsibilities.

**Challenge:** Because of existing ambiguity in defining who holds responsibility for review of data associated with publications, datasets may be approved for release without having actually been reviewed for data or metadata quality. In some cases, the specific dataset underlying an interpretive product may not be clearly identified, so an uncorrected version of the data could be inadvertently released.

**Recommendation # 7.** Datasets associated with approved publications should be held until data are reviewed and have satisfactory metadata. This recommendation is consistent with Fundamental Science Practice policy elucidated in SM 502.5 Safeguarding Unpublished U.S. Geological Survey Data, Information, and Associated Scientific Materials.

**Recommendation # 8.** To indicate underlying data for the publication have been reviewed during the approval process, DOIs should be included in both the dataset and publication metadata to clearly indicate which datasets are reviewed and approved for release.

**Challenge:** Supervisors designate metadata reviewers, who are responsible for checking the metadata associated with the data. Metadata reviewers check for compliance with standards, completeness, and accuracy in the record's description of the data. However, requirements for metadata review are inconsistent across the bureau. (Workflow Steps 5 and 7)

**Recommendation # 9.** Enforce the policy requiring metadata review by qualified reviewers for both data products and the underlying data associated with interpretive publications.

**Recommendation # 10.** Provide detailed education in the form of short training modules and downloadable resources such as checklists and exemplary metadata records for employees who are identified as Metadata Reviewers.

**Challenge:** The publication approval package is submitted to the Science Center Director, who determines whether the product is interpretive or not, and has authority to approve non-interpretive information products for release. However, widespread lack of understanding about the Science Center

Director's approval authority for release of information products often causes non-interpretive data products to be elevated to Bureau Approving Official (BAO) level for review and approval. This extends publishing timeframes and unnecessarily burdens BAOs. (Workflow Step 14)

**Recommendation # 11.** Provide guidance and standards for Science Center Directors in determining interpretive content.

**Challenge:** After the product is published, authors depend on the USGS publication system to preserve the data. However, there is no single entity within the USGS responsible for preserving USGS data, information products and associated metadata. This critical function includes curation activities such as organizing and indexing information products, services to assist in discovery, and ensuring data, metadata and publications are available. The Publications Warehouse (pubs.usgs.gov), a Web application administered by the USGS Library, manages information about, and provides access to, publications written by USGS authors. It does not store or preserve the publications themselves, but links to other systems inside and outside the USGS that may also not preserve or curate publication files and associated data. (Workflow Step 16)

**Recommendation # 12.** Establish Bureau-wide policies on data preservation, accompanied by specific guidance on best practices for preparing data for preservation at USGS. This guidance should be posted on the USGS Data Management Web site. (http://www.usgs.gov/datamanagement).

**Recommendation # 13.** Establish and support one or more curated Bureau repositories for USGS information products, data, and metadata, and require its use. It is critical to the future of USGS to keep our assets curated and connected where applicable. These official USGS repositories should be registered as designated USGS repositories and meet National Archives and Records Administration (NARA), Office of Management and Budget (OMB), Departmental, and USGS requirements. They should include properly curated and managed long-term storage capability for USGS science data, be managed and maintained by the USGS, and accommodate datasets of all sizes. Designated USGS repositories should be directly connected to the Publications Warehouse (e.g., the Publications Warehouse remains the central catalog point of entry for discovery and access of all USGS data, metadata, and publications). The repositories should be mandatory for scientists to utilize for all data, metadata, and information products, unless established, USGS Records Management Program-approved alternatives, are in place. Repository managers should work with the scientists to ensure that proper, robust metadata is provided, and that appropriate preservation and curation plans are implemented. All ongoing data curation and preservation activities should be provided by the repositories, including transfers to NARA at established times during

the life cycle of the data in accordance with record schedule requirements.

**Challenge:** USGS scientists and managers are uninformed about USGS policies and processes for release of data.

**Recommendation # 14.** Develop mandatory USGS training modules about the release of data in the USGS such that scientists and managers are better informed and understand their roles and responsibilities in the process. (for example, Departmental DOI Learn required training; short flash or video based presentations; documentation available on the Data Management Web site).

**Recommendation # 15.** Provide information about new policies and accompanying processes for releasing data through Science Center Web sites to encourage approved data release without the high cost of producing a USGS series publication.

**Challenge:** Production and release of high quality data is an activity from which many scientists do not see an immediate benefit. USGS scientists should have incentives in place to reward them for releasing high quality, documented data that adheres to good data management practices throughout the data lifecycle.

**Recommendation # 16.** Include production of high quality data products as a positive factor in the Research Grade Evaluation (RGE) process, including credit for releasing data that adheres to data management practices such that the data can be used by other USGS scientists and other scientific organizations.

**Recommendation # 17.** Adopt data citation standards so that USGS scientists receive credit for releasing data that is used by others.

# Appendix A – Team Members

--John Faundeen, Climate and Land Use Change (Sioux Falls, SD)

--Dave Ferderer, Energy and Minerals and Environmental Health (Denver, CO)

--Peter Fox, Rensselaer Polytechnic Institute (Troy, NY)

--Keith Kirk, Office of Science Quality and Integrity (Santa Cruz, CA)

--Fran Lightsom, Natural Hazards (Woods Hole, MA)

--Greg Miller, Natural Hazards (St. Petersburg, FL)

--Viv Hutchison, Core Science Systems (Denver, CO)

--Andrea Ostroff, Core Science Systems (Reston, VA)

--Stephan Zednis, Rensselaer Polytechnic Institute (Boulder, CO)

--Carolyn Reid, Office of Science Quality and Integrity (Reston, VA)

## Appendix B – Publish Data via USGS Data Series Report

**Use Case Name**

UC-1 Author and Publish Information Product using a USGS Publication Series

**Point of Contact Name**

Data Release Use Case Team point of contact for this use case: Keith Kirk, OSQI

**Goal**

Release data via publishing an information product in a USGS publication series.

**Summary**

Scenario: a USGS scientist wants to make a public release of a USGS data collection (previously not released or part of a national collection or other 'approved' dissemination).

- Scientist creates a data product in one of the USGS publication series (Data Series, Open File Report, or Administrative Report) and publishes it. Publishing involves meeting USGS requirements for:
    - review of content, metadata, policy implications, and conformity with USGS standards for accuracy and clarity of expression;
    - use of the USGS Information Product Data System (IPDS);
    - approval by the designated approving official (a Science Center Director (or designee) can approve non-interpretive data. A BAO must approve data that are considered new interpretive (refer to Survey Manual Chapter 205.18).); and
    - production and release through the USGS Science Publishing Network.
- Three publication series can be used:
    - A Data Series is currently the common USGS means of documenting data to meet approval standards.
    - An Open-File Report is intended for preliminary release of information and has minimal editorial review requirements.
    - An Administrative Report is submitted to another Federal agency that may publish, not publish, or may allow USGS to publish.
        - Note: Another series, the Techniques and Methods series is appropriate to use

for data collected and processed with new or unique techniques.

- Once published, the data is considered to be disseminated or released. The data can then be made available through value-added data services and new product formats (data.gov, Web Mapping Service (WMS), and so on).

**Actors**

- Author (primary, creates and revises product and initiates product release request)
- Supervisor (secondary, reviews for conformity with publication policies and processes)
- Peer Reviewer (secondary, reviews product for scientific quality)
- Metadata Reviewer (secondary, reviews metadata for accuracy and conformance with standards)
- Publication Editor (secondary, verifies product meets USGS standards for accuracy and clarity of expression)
- Approving Official (approves product for release)
- Curator (secondary, manages product preservation)
- Requesting Entity (primary initiator)
    - Internal Requesting Entity
    - External Requesting Entity

**User Classes**

- USGS Scientist (can act as Author, Internal Requesting Entity)
- USGS Supervisor (can act as Process Reviewer)
- BAO (can act as approving official for interpretive data)
- Science Center Director (can act as approving official for non-interpretive data because)

**Preconditions**

USGS data that will be part of the Product exists and is available to the Author.

**Triggers**

- Need for data in the form of Product to meet project or program objectives.
- Career advancement needs of the Author.
- Specific request from internal or external entity.
- Need to release data used in research, on the occasion of publishing a scientific paper.

**Basic Flow**

Scenario: a USGS scientist wants to make a public release of a USGS data collection (not previously released or part of a national collection or other 'approved' dissemination).

1. Author checks whether the data has previously been published or is associated with an approved publication [Refer to UC-2, Finding Previously Published Products.] [If found, refer to alternate flow 1.]
2. Author makes initial decision on type of Product (such as Open-File Report, Data Series). [Refer to BR1 for guidance on USGS publication series.]
3. Author develops Product.
4. Author registers Product in IPDS and recommends peer reviewers. As a result of registering the product in IPDS it is passed to the Supervisor (this is most often the author's supervisor). [In some offices, authors complete routing sheets and designated IPDS wranglers register the product in IPDS.]
5. Process Review:
    a. Supervisor verifies choice of publication series and peer reviewers, and compliance with relevant policy requirements. [BR1, BR2.]
        i. Includes informing the Science Center Director about any product content that might be sensitive or controversial and identification of any internal or external groups or agencies that might have particular and (or) immediate interest in such a product,
        ii. Determines the need for a courtesy review, and
        iii. Determines if the product meets criteria for inclusion on the USGS Peer Review Agenda, that is, should the product be considered influential science.
    b. Supervisor forwards or requests the author forward the Product to a minimum of two Peer Reviewers and one Metadata Reviewer.

6. Peer Reviewer returns Product to Author with a recommendation that Product be published and a list of recommended changes.

7. Metadata Reviewer returns Product after completing the following:

    a. Reviewer checks compliance using metadata validation tool

    b. Check that field names and data values are consistent with information in entity/attribute section

    c. Perform quality checks

        i. Check that bounding coordinates match location keywords

        ii. Check on-line linkage to data

        iii. Check that field names and data values are consistent with information in entity/attribute section

        iv. Check that field names and data values are consistent with information in entity/attribute section

8. Decision point: do reviewers recommend major revision of product?

    a. If yes author makes major revision and returns to step 5.

    b. If no author proceeds to step 9.

9. Author revises Product in response to Reviewer recommendations and documents reconciliation and responses and sends revised Product to Publication Editor.

    a. reconciliation reports how review comments were addressed

10. Publication Editor reviews Product for compliance with USGS standards for accuracy and clarity of expression and returns to Author with a list of recommended changes. [BR3.]

11. Author revises Product in response to Editor recommendations.

12. Author places relevant materials in IPDS document vault and forwards to Supervisor. [In some offices, Author compiles all versions of product and related documentation and an IPDS wrangler creates the request.]

    a. These materials include:

        i. the original manuscript,

ii the revised manuscript in response to peer and editorial review,

iii all original peer reviewers' comments, including memoranda or emails from reviewers and any manuscript markups, and

iv reconciliation document addressing peer review comments.

13 Supervisor verifies that publication request is complete and responses/reconciliations are appropriate. Supervisor forwards request to local Approving Official (Science Center Director).

14 Local Approving Official determines whether Product is interpretive or not, and forwards request to BAO for approval if product is interpretive. [BR5, BR6.]

15 Approving Official reviews Product publication request, and approves request. [BR4.]

16 Product is preserved by Publication Warehouse.

17 Product is published by Science Publishing Network.

18 Data is approved for release.

***Alternate Flows***

1 If an existing product is found, Author releases data to requesting entity and adds Product to CV if appropriate.

**Post Conditions**

- Product produced, published, and preserved.
- Product deposited in Publications Warehouse.
- Metadata is preserved as part of Product, but might not be deposited in a clearinghouse.
- Documentation of Product review and approval is preserved in the IPDS or in a file system at originating Science Center.

***Activity Diagram***

- Refer to Figure 1.

**Assumptions**

- USGS data for information product is assumed to be non-proprietary, and non-sensitive

**Business Rules**

- BR1. Guidance on USGS publication series is at http://www.usgs.gov/usgs-manual/1100/1100-3.html.
- BR2. Policy about technical review is at http://www.usgs.gov/usgs-manual/500/502-3.html.
- BR3. Policy about editorial review is at http://www.usgs.gov/usgs-manual/1100/1100-2.html. In addition, ***Suggestions to Authors of the Reports of the United States Geological Survey*** is online at http://internal.usgs.gov/publishing/sta/.
- BR4. Policy about publication approval is at http://www.usgs.gov/usgs-manual/500/502-4.html.
- BR5. Policy on approval authorities is at http://www.usgs.gov/usgs-manual/200/205-18.html.
- BR6. Clarification about "interpretive" data products is available at the Fundamental Science Practices FAQ, http://internal.usgs.gov/fsp/faqs.html#def6 .
- Additional USGS policies are available through the Fundamental Science Practices Web site, http://internal.usgs.gov/fsp/policies.html.

**Notes**

- Related Use Cases:
    - UC-2, Finding Previously Published Products
- Other options:
    - Provisional release with disclaimer statement: with the understanding that the final product may contain revisions. What is the desirable lifetime of these forms? At present there are no formal requirements for metadata generation (also provisional). Discretion on release decision can be made by the Science Center.
    - Release through approved publication using an external publisher instead of a USGS publication series.
    - Release on USGS Science Center Web site (UC-3 *under development*)