USGS
science for a changing world

A Roadmap for Enabling Integrated Science: The USGS Experience in Open Data

Twentieth International Conference on Grey Literature
"Research Data Fuels and Sustains Grey Literature"

Viv Hutchison
US Geological Survey
Science Analytics & Synthesis (SAS)

December, 2018

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

USGS
science for a changing world

Viv Hutchison

- USGS employee since 2002
- Background in Library and Information Science (MLS from University of Maryland)
- Current Position(s):
  - Science Data Management Branch Chief
  - Acting USGS Library Director

USGS
science for a changing world

Topics

USGS Science Data

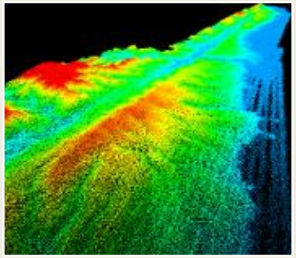
USGS Science Publications

USGS Path to Open Science

US Geological Survey

Science for a Changing World

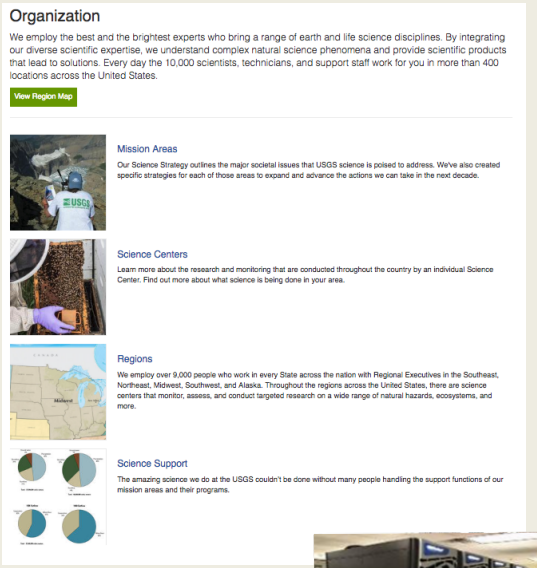
The USGS serves the Nation by providing reliable scientific information to describe and understand the Earth;
minimizing loss of life and property from natural disasters;
managing water, biological, energy, and mineral resources; and enhancing and protecting our quality of life.



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

US Geological Survey: Organization

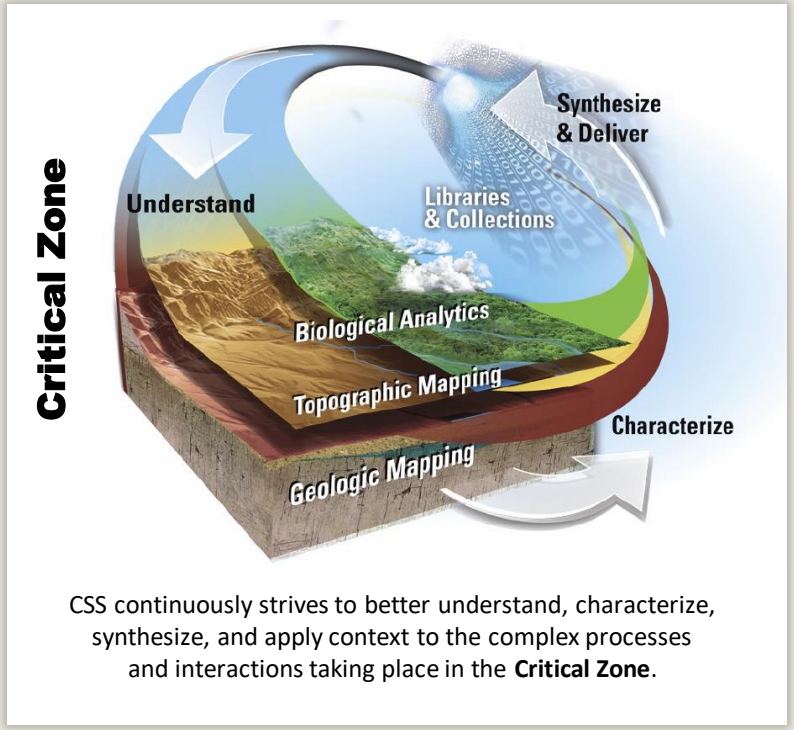
- 7 Mission Areas
  - Core Science Systems
  - Ecosystems
  - Energy and Minerals
  - Environmental Health
  - Land Resources
  - Natural Hazards
  - Water
- 7 Regions
- Over 90 Science Centers + Research Stations (totaling over 400 facilities)
- Science Support Offices



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

USGS
science for a changing world

US Geological Survey
Core Science Systems – *Mission Overview*



CSS continuously strives to better understand, characterize, synthesize, and apply context to the complex processes and interactions taking place in the **Critical Zone**.

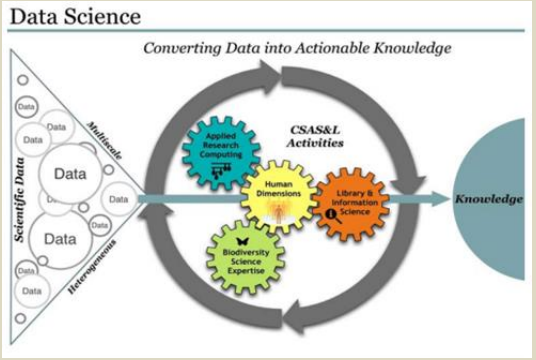
America relies on our broad range of nationally-focused Earth system science, geospatial research, and foundational data to remain informed, safe, and healthy.

- **Provide baseline geospatial, geological, and biogeographical data** for the Nation's topography, natural landscape, and human-constructed environment.
- **Conduct science, surveys, and research** on the Nation's geological and biological resources and offer data, map services, and decision-support tools for the public to easily discover/use for local, regional, large-scale, national, or continental analyses.
- **Coordinate effective and economical use and management of geospatial data assets** for the government and the Nation.
- **Conduct data synthesis and analysis across disciplines** to enable data driven science in support of energy and mineral assessments, infrastructure modernization, natural hazards mitigation, land and water management and conservation, and national security.
- **Connect USGS scientists to published literature, data, and collections** to support evidence-based practices.
- **Inventory and preserve geological and geophysical data collections** to provide a framework for geoscience data and information sharing.
- **Improve Federal-State cooperation and collaboration** by effectively leveraging partnerships.
- **Provide computational capabilities** through high—performance computing infrastructure, workshops, and expertise.

Science Analytics & Synthesis (SAS)

SAS Mission

- **Accelerate research and decision making through data science, information delivery, advanced computing and biodiversity analytics.**
- **Key Science Themes**
  - **Data Science**
  - **Biodiversity Science**
  - **Advanced Research Computing**
  - **USGS Library**
  - **Science Data Management**

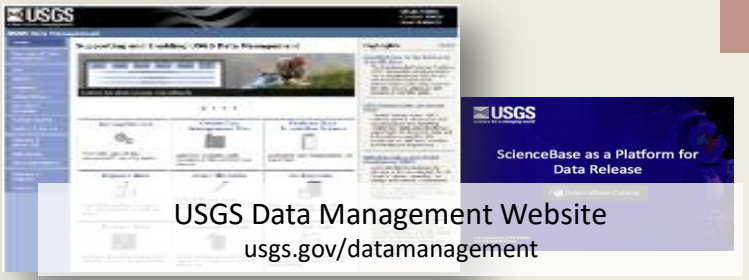


SAS Science Data Management Branch

providing tools, best practices, and education for data management

Policies

Education



USGS Data Management Website
usgs.gov/datamanagement

Community for Data Integration
usgs.gov/cdi

Partnerships

USGS Powell Center

CSIP

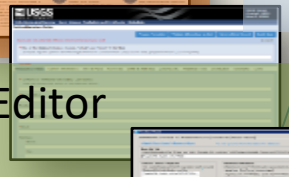
DataONE

Science Data Tools

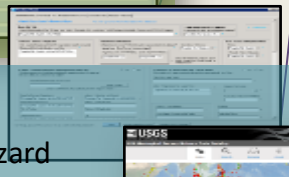
ScienceBase
sciencebase.gov

Data Release Workbench
https://data.usgs.gov/datarelease/

DOI Creation Tool
www1.usgs.gov/csas/doi

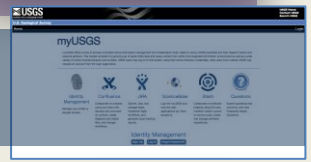


Online Metadata Editor
www1.usgs.gov/csas/ome



Metadata Wizard
sciencebase.gov/metadatawizard

Science Data Catalog
data.usgs.gov

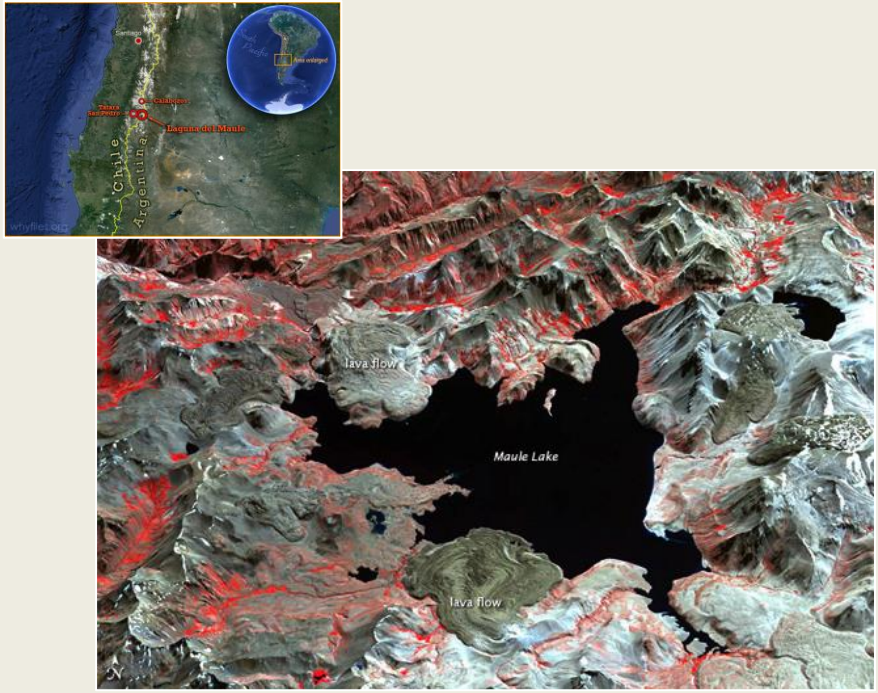


myUSGS
my.usgs.gov/resources

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

USGS
science for a changing world

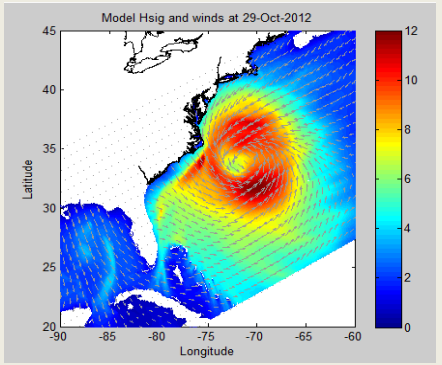
USGS Science Data



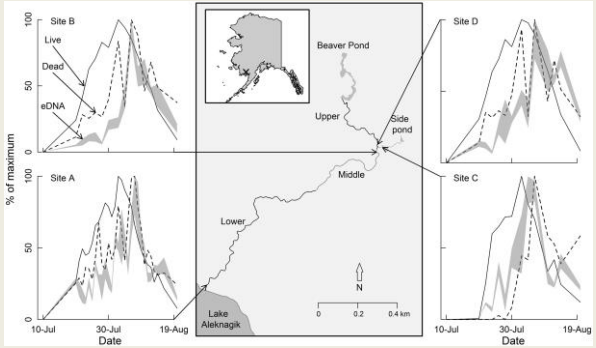
U.S. Department of the Interior
U.S. Geological Survey

USGS Science Data

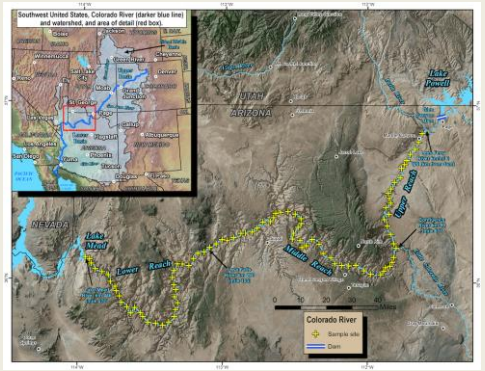
- Multiple disciplines of Earth and biological sciences
- Distributed environment in many Science Centers
- Labs, field collection, instrument collection, mapping, remote sensing, etc
- Many data formats
- Wide variety of data sizes
- Data integration needs



Modeling wave heights due to surface winds in Hurricane Sandy



Hansen Creek, Alaska showing locations of eDNA sampling sites and visual count survey reaches



Large River monitoring & mapping

USGS
science for a changing world

Example of USGS science data in action

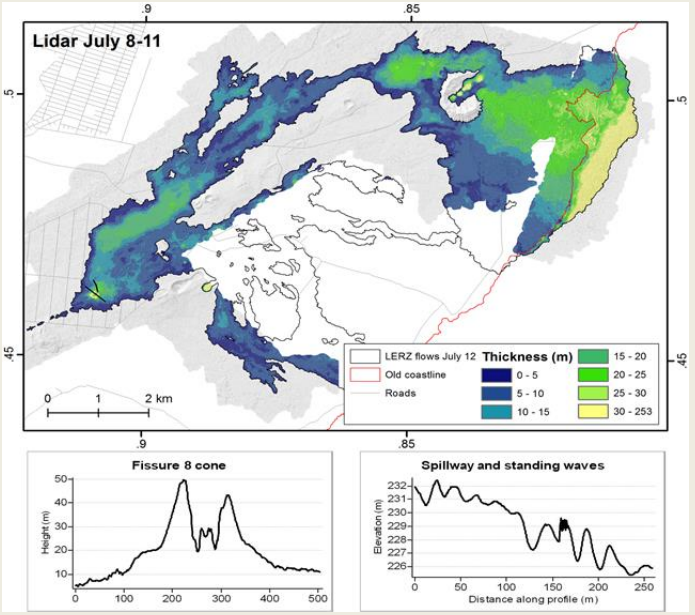
U.S. Department of the Interior
U.S. Geological Survey

The Event(s)

- Lava flows travelling at 3 ft/s (2 mph)
- Ash plumes to 10,000 ft from main vent
- 24 fissures, 19 active (mid August 2018)
- Major Damage to Leilani Estates
- Encroached on a hydrothermal power plant

Fissure 8 - May 28, 2018

Rapidly producing new Coastline



Damage in Volcanoes NP

Before…

After…

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

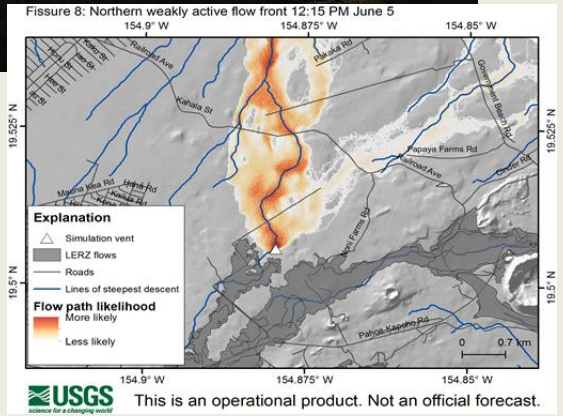
Lava Flow Hazards & Damage

Hannah Dietterich, USGS Volcano Hazards Program, Alaska Volcano Observatory

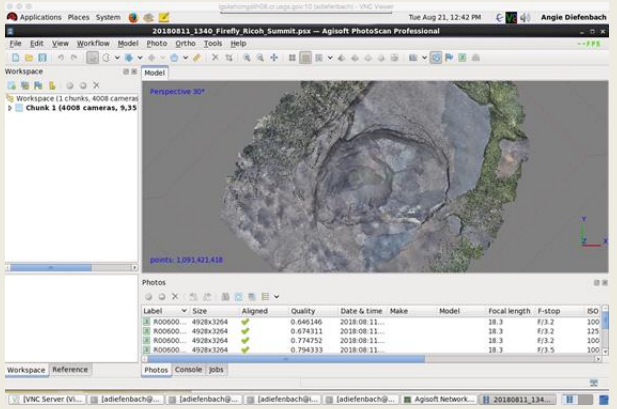
- Tracking new vent formation on Kilauea
- Estimating potential damage impacts due to lava flow and inundation
- Multiple-agencies involved - FEMA, USGS, BLM, NPS, NASA

Rapid data analysis using:

- Lidar and Unmanned Aerial Systems (UAS) data collected to update Digital Elevation Models (DEM)
- High Performance Computing



Lava Flow Hazards & Damage

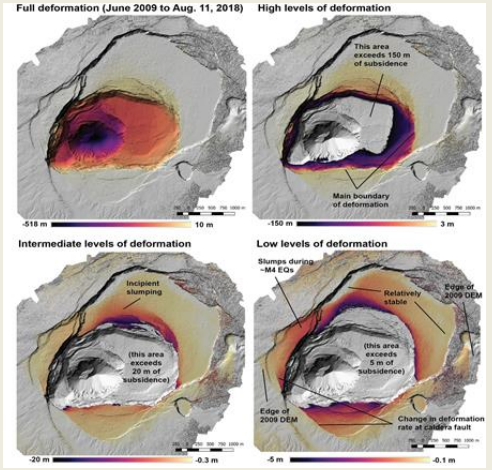


Angie Diefenbach, USGS Cascades Volcano Observatory

- Developed models based on imagery captured with Ricoh GII camera using Firefly Pro Vertical Take off and Landing (VTOL) fixed wing UAS
- 4,008 images (25GB) from UAV mounted camera generated several "products"
- Dense Point Cloud with 1,091,421,418 points (27GB)
- Orthomosaic (2.1GB)
- DEM (3.0GB)

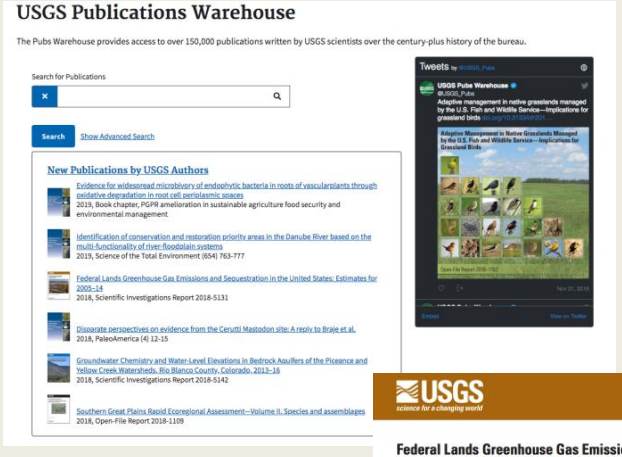
Data products used to understand the extent of the caldera collapse at the summit that occurred during the 2018 eruption

- As much as -518m change in elevation (slumping) has been observed at the summit



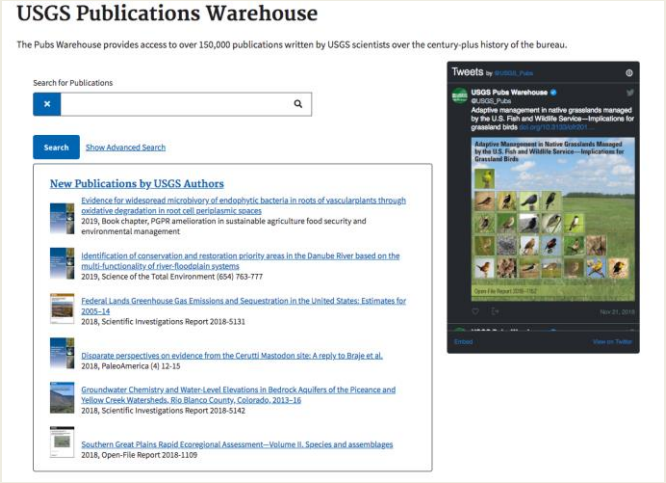
USGS
science for a changing world

USGS Science Publications



USGS Publications Warehouse

- Authoritative resource and public interface to USGS authored publications
- Provides metadata and links to USGS funded and authored publications
- USGS Library leads Pubs Warehouse



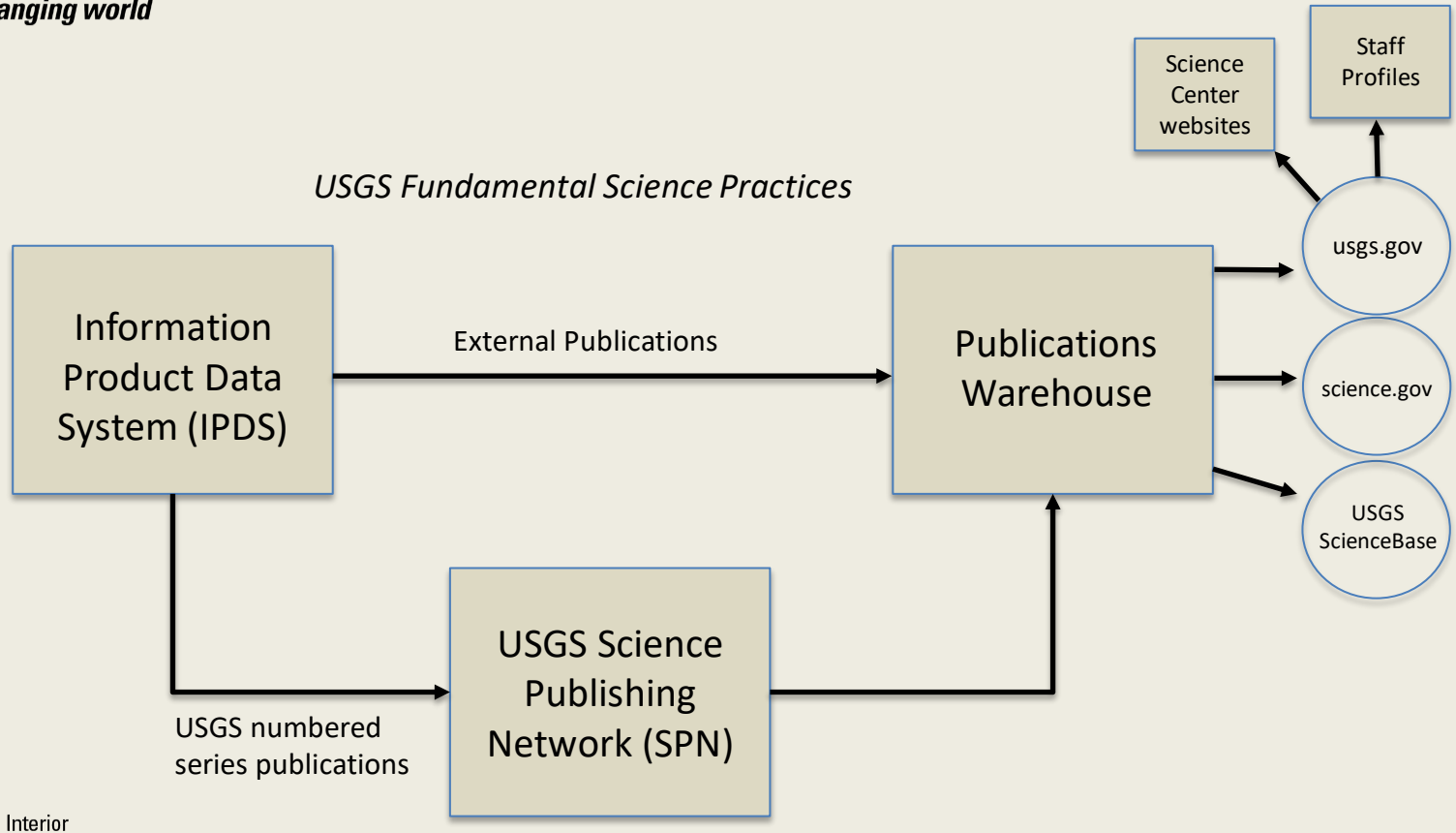
https://pubs.er.usgs.gov

USGS
science for a changing world

Publications Warehouse Content

Published by USGS	Published by external entity
• USGS numbered series reports • Other USGS reports	• Journal articles • Conference proceedings • Books • Book chapters • Other government reports

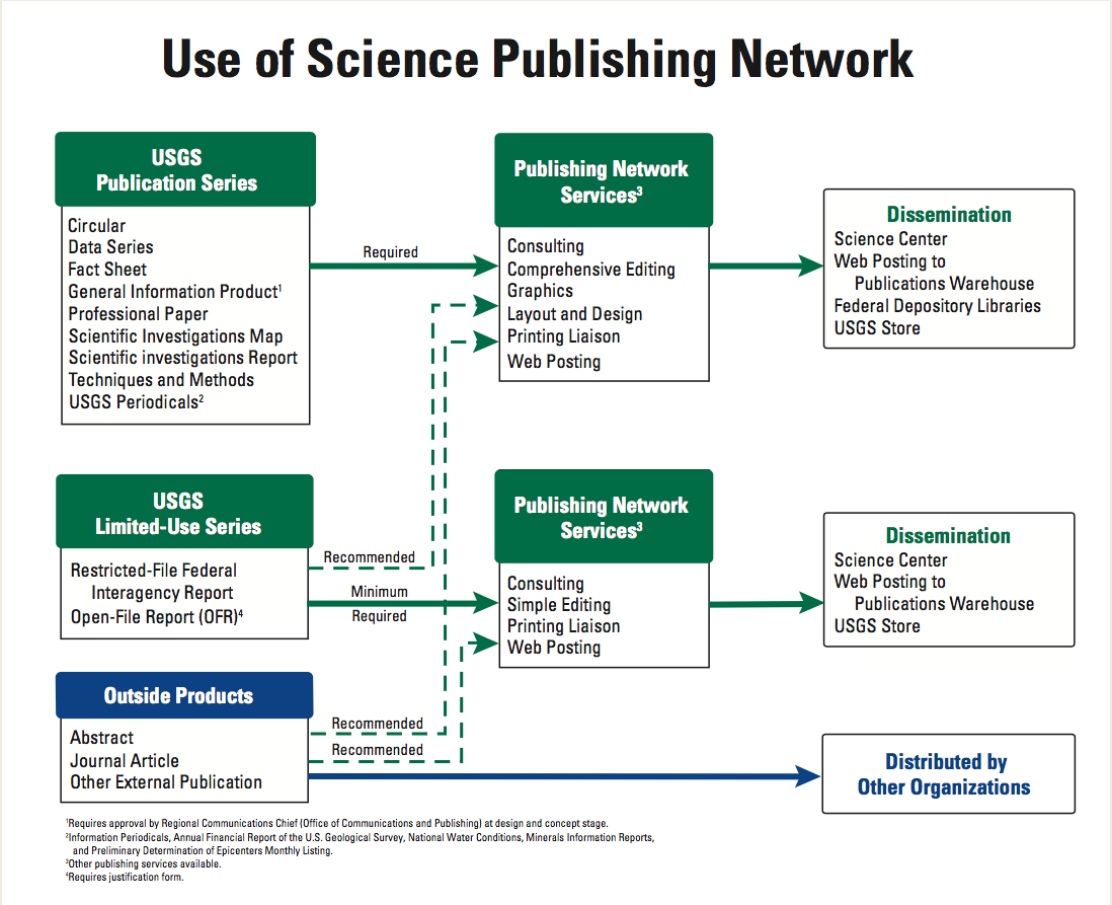
Publications Workflow



Science Publishing Network: Services

Science Publishing Network (SPN)

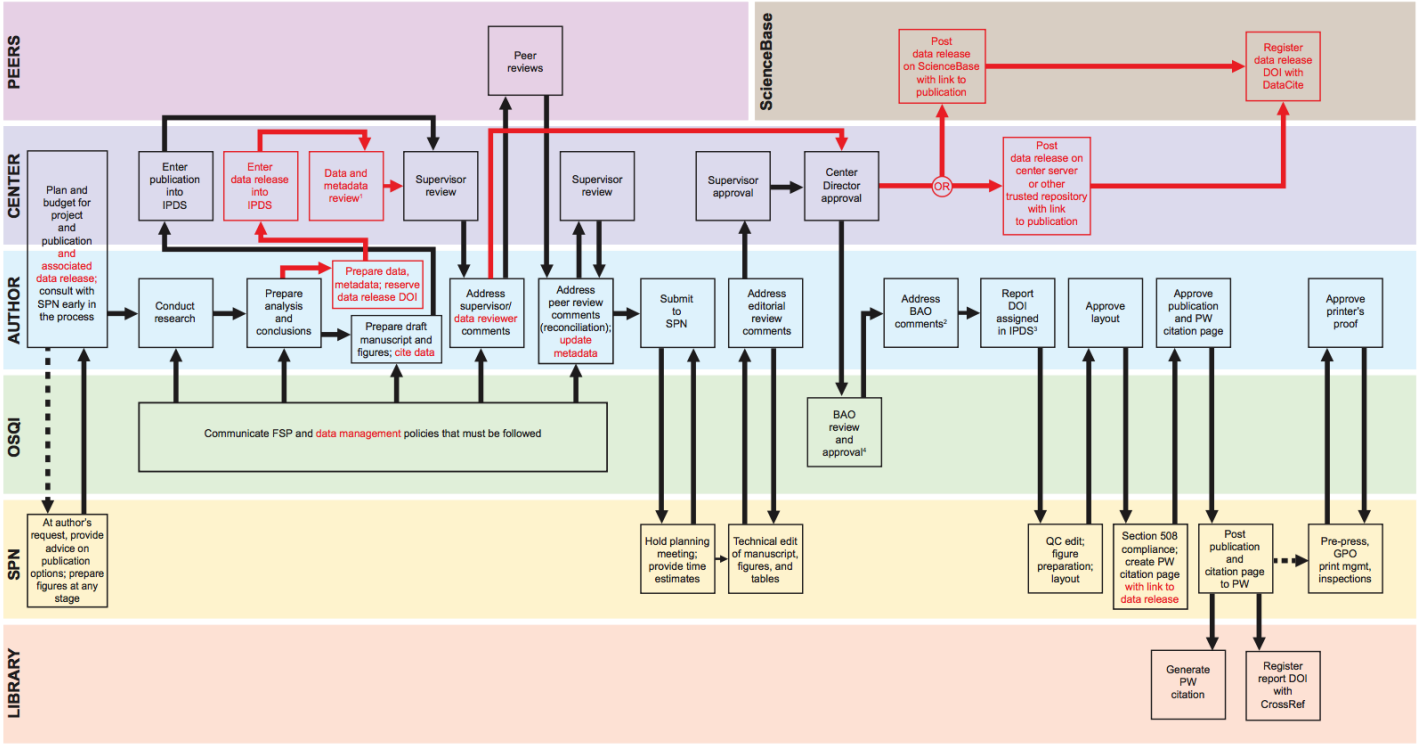
- Product planning and consultation
- Publication policy advising
- Map editing and production
- Scientific text editing
- Illustration
- Graphic design and layout
- Online publication support



Publications + Data Release Process for Authors

Overview of the Publishing Process for USGS Series Publications Accompanied by a Data Release

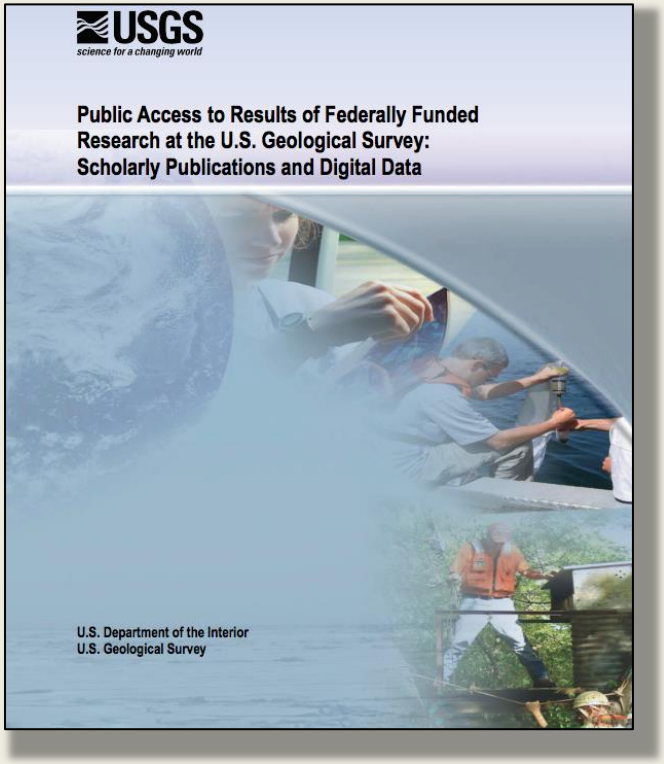
[Any dataset used to support the conclusions in a series publication must be released simultaneously with the publication, unless the data have been previously released. Abbreviations: BAO, Bureau Approving Official; DOI, Digital Object Identifier; FSP, Fundamental Science Practices; GPO, Government Publishing Office; IPDS, Information Product Data System; mgmt, management; OSQI, Office of Science Quality and Integrity; PDF, Portable Document Format; PW, Publications Warehouse; QC, quality control; SPN, Science Publishing Network; USGS, U.S. Geological Survey]



[1]See Survey Manual chapter 502.8 (https://www2.usgs.gov/usgs-manual/500/502-8.html).
[2]For an administrative report, this workflow is followed to the "Address BAO comments" step. Administrative reports document studies that are released to other Federal agencies. Administrative reports are considered to be unpublished.
[3]The report DOI is assigned automatically in IPDS after the author approves the "Prepare for SPN Production" step.
[4]Only interpretive products go to the BAO; noninterpretive products are granted Bureau approval by center directors or their designees.

USGS
science for a changing world

USGS Path to Open Science



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

Drivers: Government Initiatives

Broader Ecosystem of Government Mandates:

- Open Government Initiative (2009)
  - transparency, participation, and collaboration
- Digital Government Strategy (2012)
  - "Digital Government: Building a 21st Century Platform to Better Serve the American People"
- Open Data Initiatives (2013)

USGS
science for a changing world

Open Data Initiatives – 2013

Executive Order

- Increasing Public Access to the Results of Federally Funded Scientific Research Feb. 22, 2013

Executive Order

- Making Data Open and Machine Readable: the New Default for Government Information May 9, 2013

M-13-13 Open Data Policy

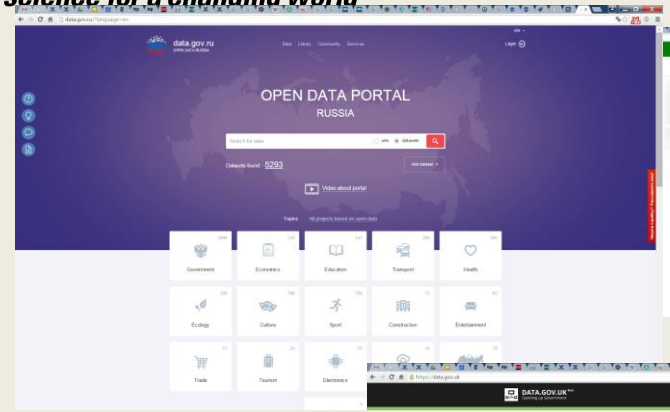
- Managing Information as an Asset

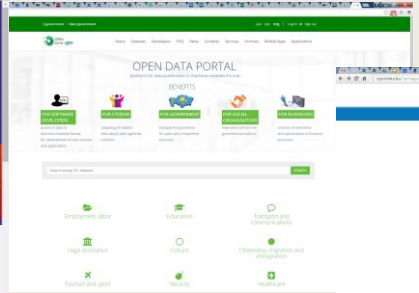
Highlights:

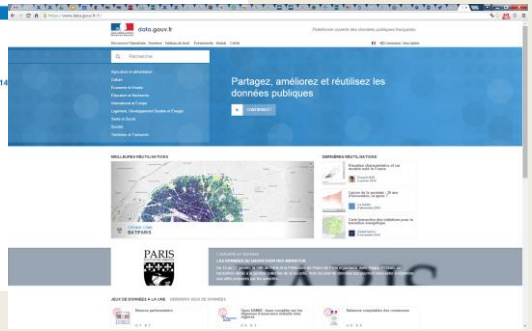
- Data must be **accessible** (in open formats)
- Data must be **described**
- Data must be **reusable**
- **Complete** (published in "primary form" with finest level of granularity)
- **Timely**
- **Managed** post-release

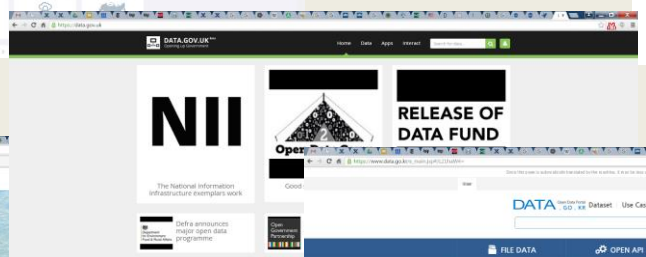
USGS
science for a changing world

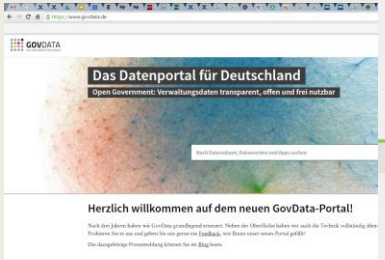
Becoming a Global Open Data Movement: Governments around the world

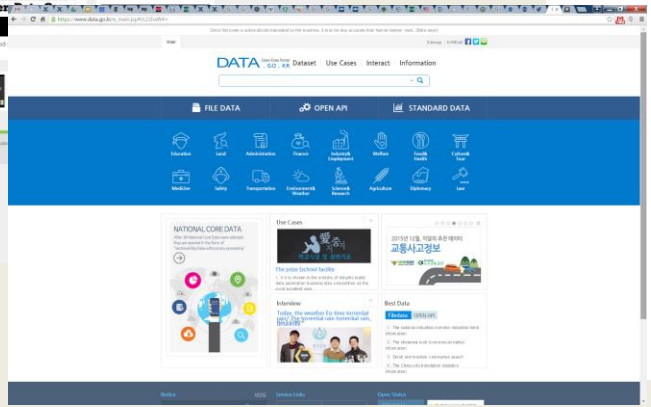


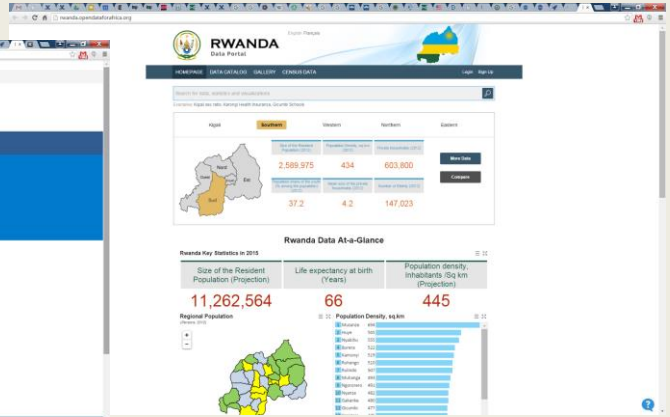










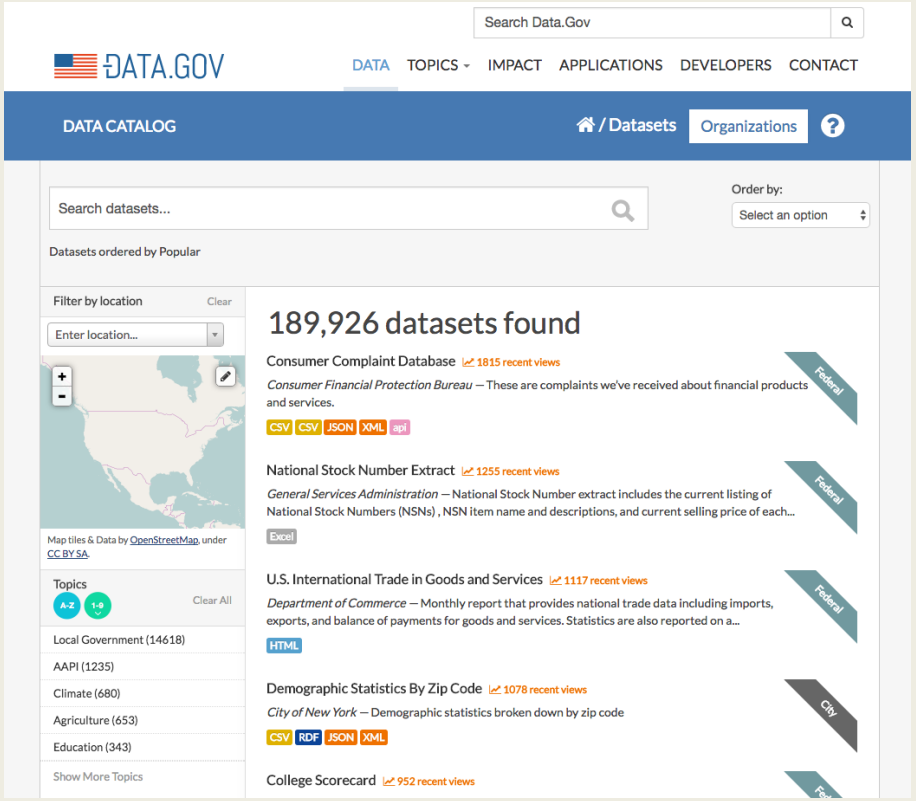


Department of the Interior: Open Data Coordination Efforts

Coordinated Interior metadata contributions to data.gov

Sponsored Cross-Agency Working Groups:

- Data Governance Board
- Data Services Team
- DOI policies
- Information Management & Technology Strategic Priorities Groups (e.g., “Managing Data as a Strategic Asset”; “Risk Based Portfolio Management”…)

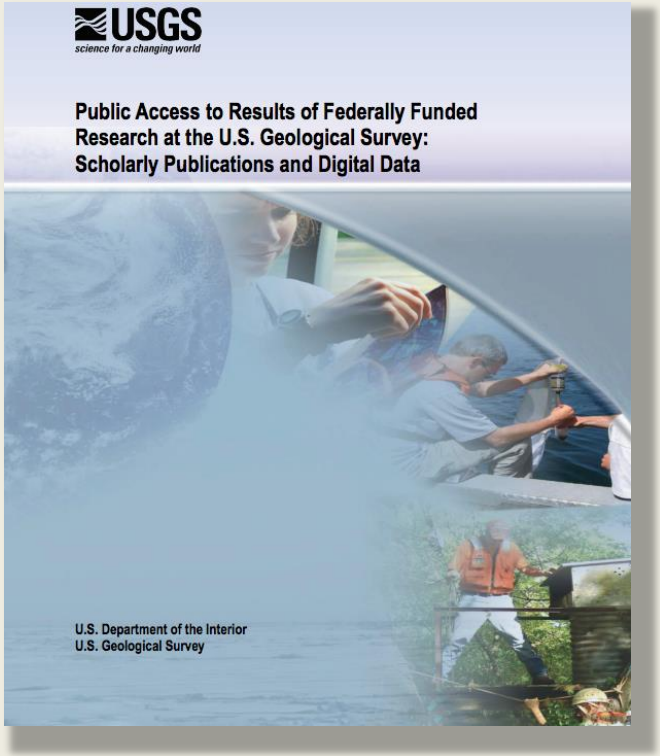


USGS Publishes Public Access Plan

M-13-13 Open Data Policy

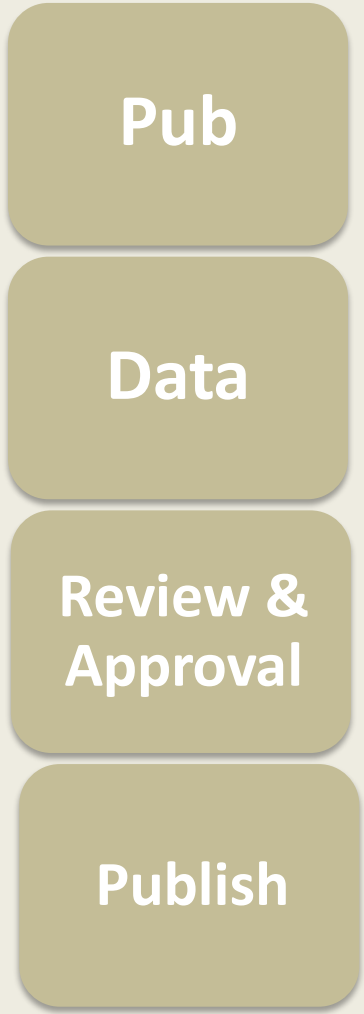
USGS Response, published in 2016:

"Public Access to Results of Federally Funded Research at the U.S. Geological Survey: Scholarly Publications and Digital Data"



USGS Public Access Plan

- Regarding data, the Plan introduces a “new normal”:
  - USGS scientists must release the data upon which their scientific publications are based.
- This requires, for data:
  - A detailed **data management plan** for all research projects
  - Data in machine readable, **open formats**
  - Standard **metadata** describing the data
  - A **digital object identifier** for the data, recorded in the metadata record
  - **Approvals**: requiring data review, metadata review, and either Center Director or Bureau Approving Officials
  - Hosted and shared from a reliable, **repository** location
  - Metadata **shared** through the USGS Science Data Catalog



Implementation of the Public Access Plan in the USGS

- The “new normal” was a surprise to some.
- It requires a new workflow for scientists.
- It represents a culture shift for USGS.

Building on Foundations

Policies

Fundamental Science Practices

Applications

Data and Code Management Tools

Publications Warehouse

Review and Approval Tools

People

Bureau Approving Officials (BAOs)

Science Publishing Network (SPN)

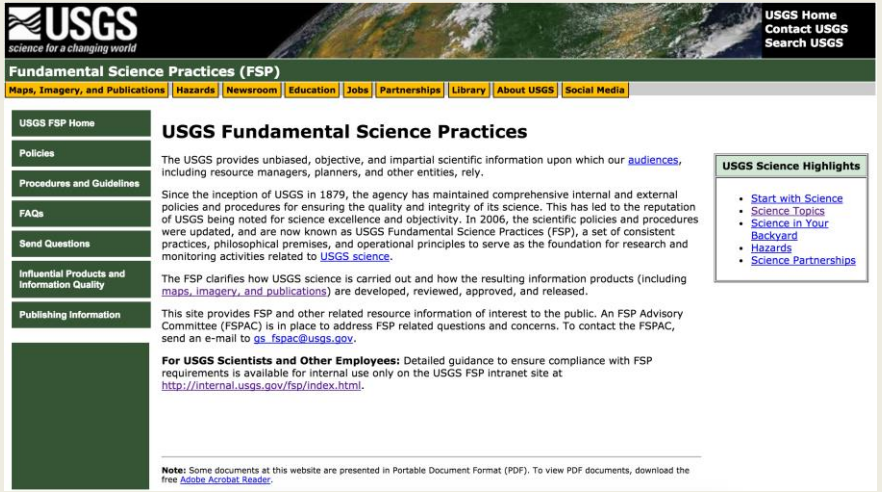
Community for Data Integration (CDI)

SAS Science Data Management

Fundamental Science Practices (FSP)

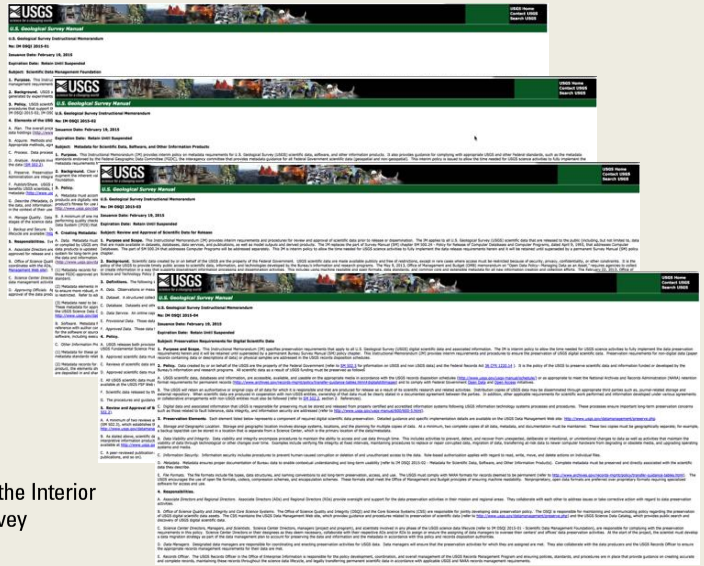
Fundamental Science Practices

- Established 2006
- FSP is a set of consistent practices, philosophical premises, and operational principles to serve as the foundation for research and monitoring activities related to USGS science
- FSPAC: FSP Advisory Council provides guidance/direction for FSP implementation



FSP includes USGS Policies

Fundamental Science Practices



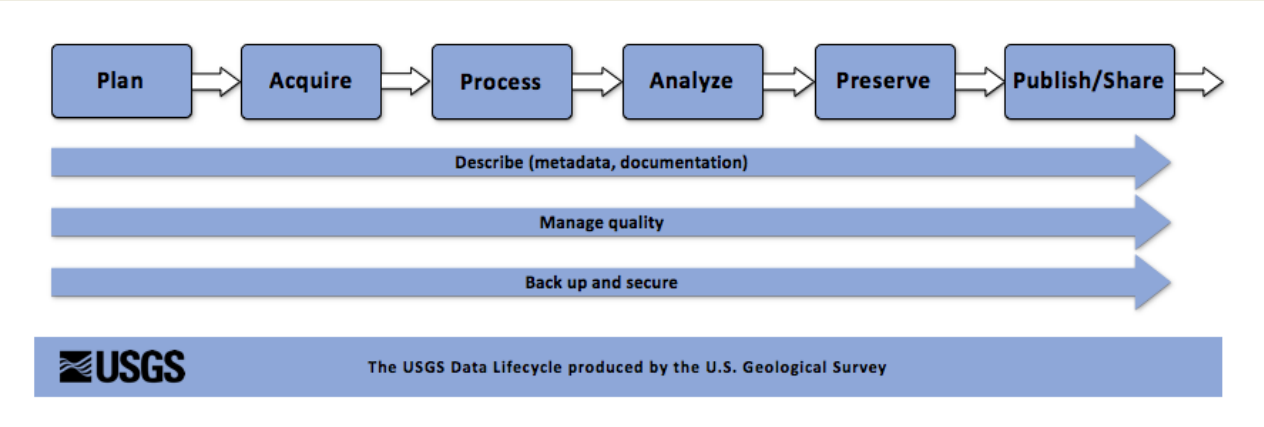
USGS released 4 new policies for data management & public access

- Scientific Data Management Foundation *(requires DMPs)*
- Metadata for Scientific Data, Software, and Other Information Products
- Review and Approval of Scientific Data for Release
- Preservation Requirements for Digital Scientific Data

USGS Science Data Lifecycle Model

Fundamental Science Practices

- The new data policies and accompanying tools are based on the model



Bureau Approving Officials (BAOs)

Bureau Approving Officials

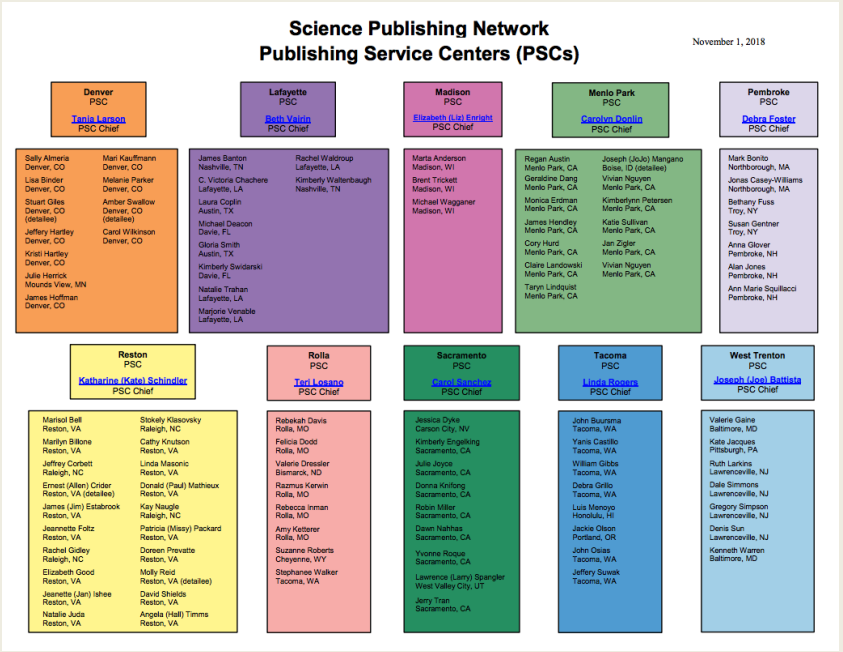
- Team of science experts
- Office of Science Quality and Integrity
- Provide approval for release of interpretive science information products including publications, data, and others

Approval levels for information products prepared by USGS authors			
Information Product Type	**Description**	**Bureau Approval**	
		Science Center Director or equivalent	**BAO in OSQI**
Circular	Synthesis based on approved published products. This series should not be used to present new data or new scientific findings.		X
Data Series	Data, multimedia, and (or) noninterpretive.	X	
Open-File Report	Interpretive and preliminary or interim release pending formal series release.		X
	Data, multimedia, and (or) noninterpretive.	X	
Administrative Report	Data (studies and findings) and (or) noninterpretive; proprietary interest to Federal agency cooperator.	X	
	Interpretive or policy sensitive; proprietary interest to Federal agency cooperator.		X
Fact Sheet	Interpretive and based on approved published products, not policy sensitive.	X	
	Policy sensitive.		X
	Noninterpretive and (or) programmatic.	X	

Science Publishing Network (SPN)

Science Publishing Network

- Distributed team of editors and publishing specialists for USGS
- Oversee internal publishing process for USGS scientists
- Coordinate with the Publications Warehouse team



Community for Data Integration (CDI)

Community for Data Integration (CDI)

- Purpose:
  - Advance understanding of Earth systems through enhanced use of data and information
  - Provide a forum for sharing ideas and learning new skills
  - Increase the visibility of work with data and information
- Over 1,000 members
- Monthly Meetings
- Open to all (inside and outside USGS)
- Working Groups
  - Data Management Working Group, seeking to elevate the perceived value of science data management

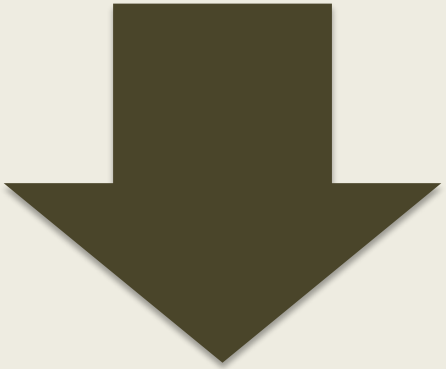
Executive Sponsors: Kevin Gallagher and Tim Quinn

CDI Data Management Working Group

Communities of practice are groups of people who share a concern or a passion for something they do and learn how to do it better as they interact regularly – Etienne Wenger, 2011

Balancing the journey to publicly accessible data

Fundamental Science Practices



USGS Policies, Open Data Initiatives, Public Access Plan

(top down)

Community for Data Integration (CDI)

Community for Data Integration (CDI)

Best Practices, Processes & Tools

(bottom up)

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

Building on Foundations: Tools and Applications

USGS and ORCiDs

- Open Researcher and Contributor ID
- Unique identifier for people
- Helps with name disambiguation
- Per Public Access Plan, now required for all publishing authors

ORCID
Connecting Research and Researchers

ORCID as a URL

http://orcid.org/0000-0003-2609-364X

Eighteen years (1996-2014) of channel cross-sectional measurements made in Spring Creek after the 1996 Buffalo Creek wildfire and subsequent flood

John A. Moody, Deborah Martin

2017, Dataset

The consequence of the 1996 Buffalo Creek wildfire disturbance and a subsequent high-intensity summer convective rain storm (~100 mm h-1) was the deposition of a sediment superslug in the Spring Creek basin (26.8 km2) of the Front Range Mountains in Colorado. Changes in the superslug near the confluence of Spring...

Synthesis of soil-hydraulic properties and infiltration timescales in wildfire-affected soils

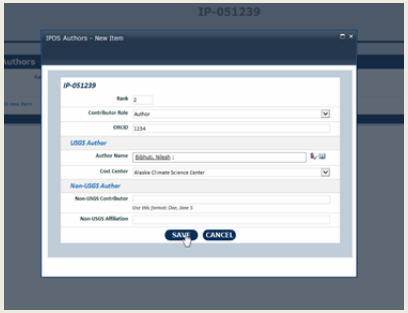
Brian A. Ebel John A. Moody

2017, Hydrological Processes (31) 324-340

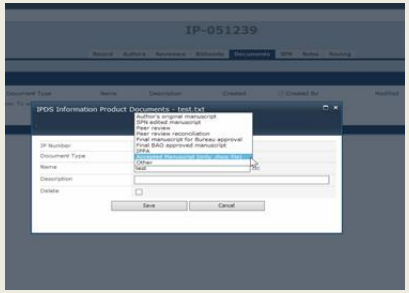
We collected soil-hydraulic property data from the literature for wildfire-affected soils, ash, and unburned soils. These data were used to calculate metrics and timescales of hydrologic response related to infiltration and surface runoff generation. Sorptivity (S) and wetting front potential (Ψf) were significantly different (lower) in burned soils compared with...

Information Product Data System (IPDS)

- Internal to USGS
- Authors, BAOs, Center Directors use IPDS to track publications;
- Per Public Access Plan, updated to include:
  - data release review and approval capability
  - dark archive for manuscripts



ORCiD field

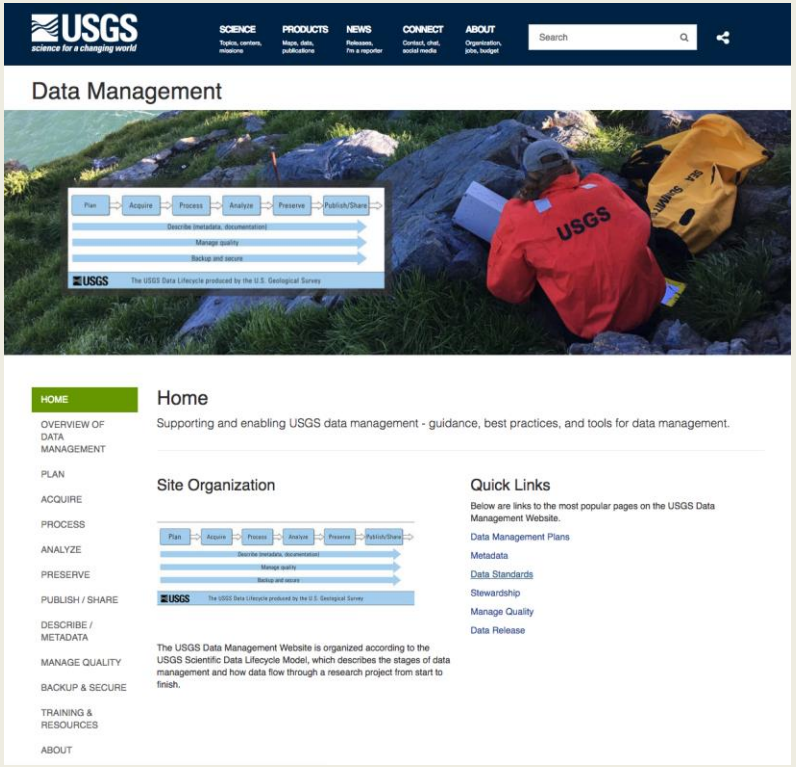


Accepted Manuscript field

USGS Data Management Website

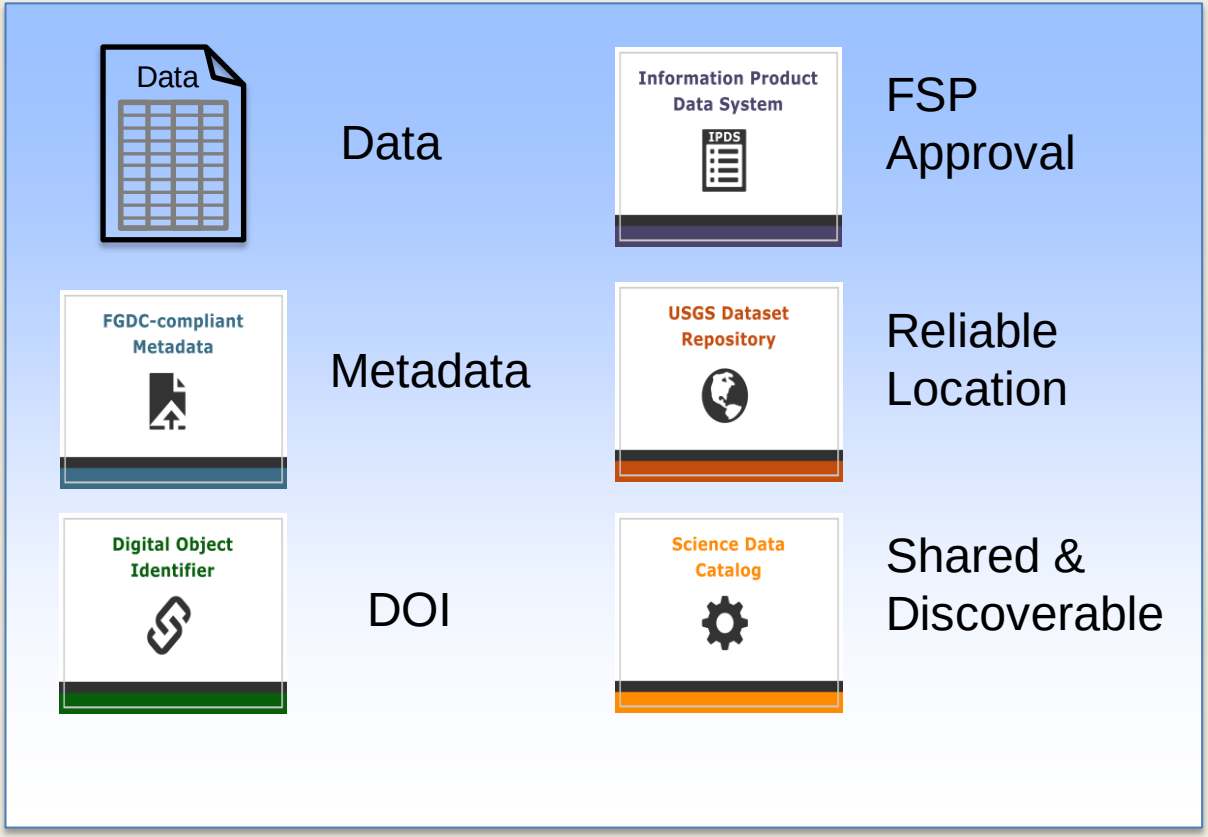
- Designed around science data lifecycle model

  Best practices, tools, links to policy



https://www.usgs.gov/products/data-and-tools/data-management

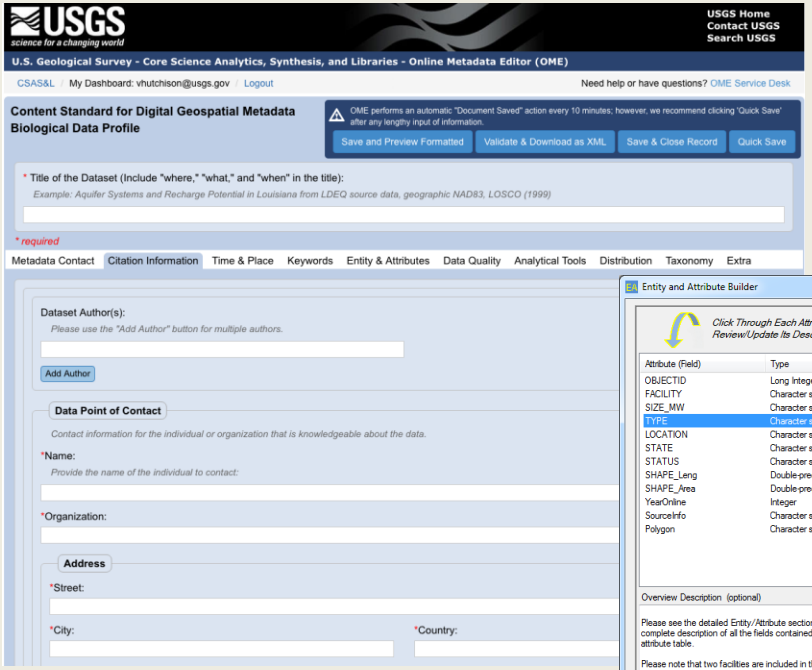
Public Access Plan: Requirements for Data



USGS Metadata Tools

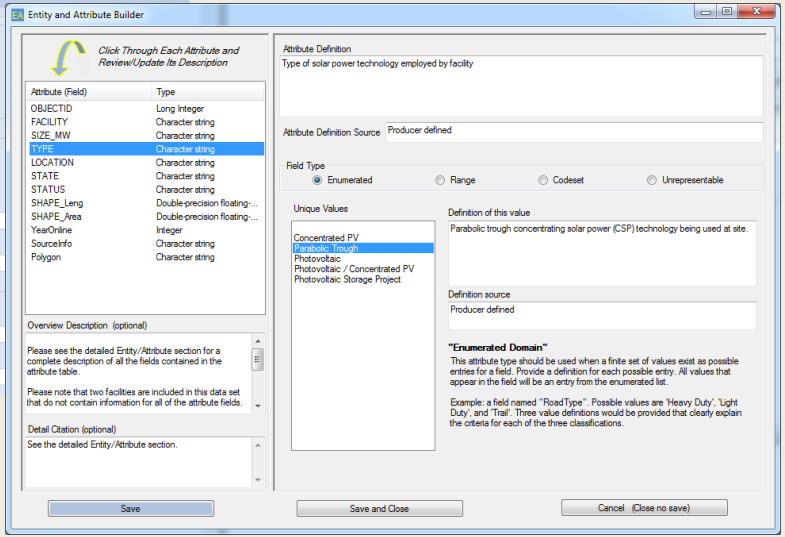
Online Metadata Editor (OME)

- Allows scientists to create and manage metadata records that describe their datasets
- Uses common questions about the data
- Based on FGDC standard (something hidden from scientists)
- Validate and download records
- (internal to USGS)



Metadata Wizard

- Plug-in tool for ESRI ArcGIS tools
- Allows scientists to create metadata records that describe their geospatial datasets
- Based on FGDC standard
- Publicly accessible

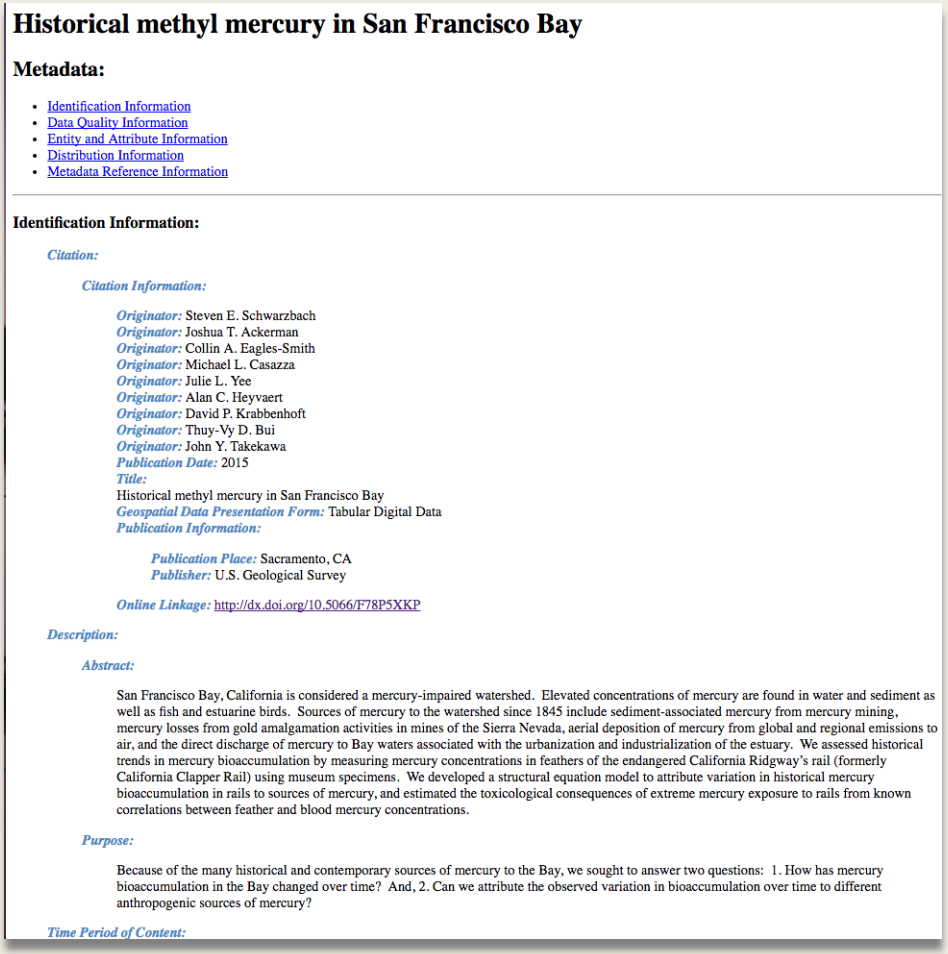


sciencebase.gov/metadatawizard

USGS
science for a changing world

Robust metadata records are produced using OME and/or Wizard:

- Metadata describes data: who produced it, how (what processes were used), data quality testing, entities and attributes, etc
- Metadata records include digital object identifiers



Historical methyl mercury in San Francisco Bay

Metadata:

Identification Information:

Citation:

Citation Information:

Originator: Steven E. Schwarzbach
Originator: Joshua T. Ackerman
Originator: Collin A. Eagles-Smith
Originator: Michael L. Casazza
Originator: Julie L. Yee
Originator: Alan C. Heyvaert
Originator: David P. Krabbenhoft
Originator: Thuy-Vy D. Bui
Originator: John Y. Takekawa
Publication Date: 2015
Title:
Historical methyl mercury in San Francisco Bay
Geospatial Data Presentation Form: Tabular Digital Data
Publication Information:

Publication Place: Sacramento, CA
Publisher: U.S. Geological Survey

Online Linkage: http://dx.doi.org/10.5066/F78P5XKP

Description:

Abstract:

San Francisco Bay, California is considered a mercury-impaired watershed. Elevated concentrations of mercury are found in water and sediment as well as fish and estuarine birds. Sources of mercury to the watershed since 1845 include sediment-associated mercury from mercury mining, mercury losses from gold amalgamation activities in mines of the Sierra Nevada, aerial deposition of mercury from global and regional emissions to air, and the direct discharge of mercury to Bay waters associated with the urbanization and industrialization of the estuary. We assessed historical trends in mercury bioaccumulation by measuring mercury concentrations in feathers of the endangered California Ridgway's rail (formerly California Clapper Rail) using museum specimens. We developed a structural equation model to attribute variation in historical mercury bioaccumulation in rails to sources of mercury, and estimated the toxicological consequences of extreme mercury exposure to rails from known correlations between feather and blood mercury concentrations.

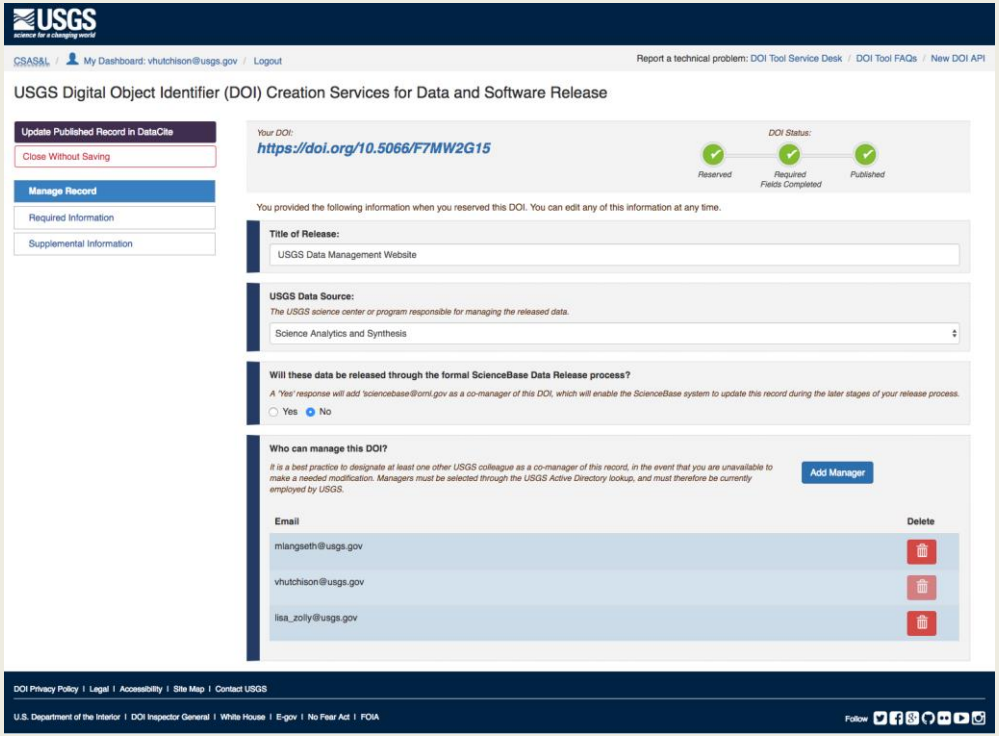
Purpose:

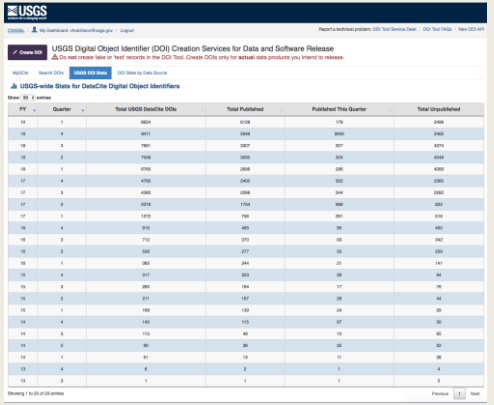
Because of the many historical and contemporary sources of mercury to the Bay, we sought to answer two questions: 1. How has mercury bioaccumulation in the Bay changed over time? And, 2. Can we attribute the observed variation in bioaccumulation over time to different anthropogenic sources of mercury?

Time Period of Content:

USGS Digital Object Identifier (DOI) Tool

- Allows creation of an identifier for datasets
- Connection to DataCite
- Identifiers for USGS data only
- DOIs are used in citation and in metadata records





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

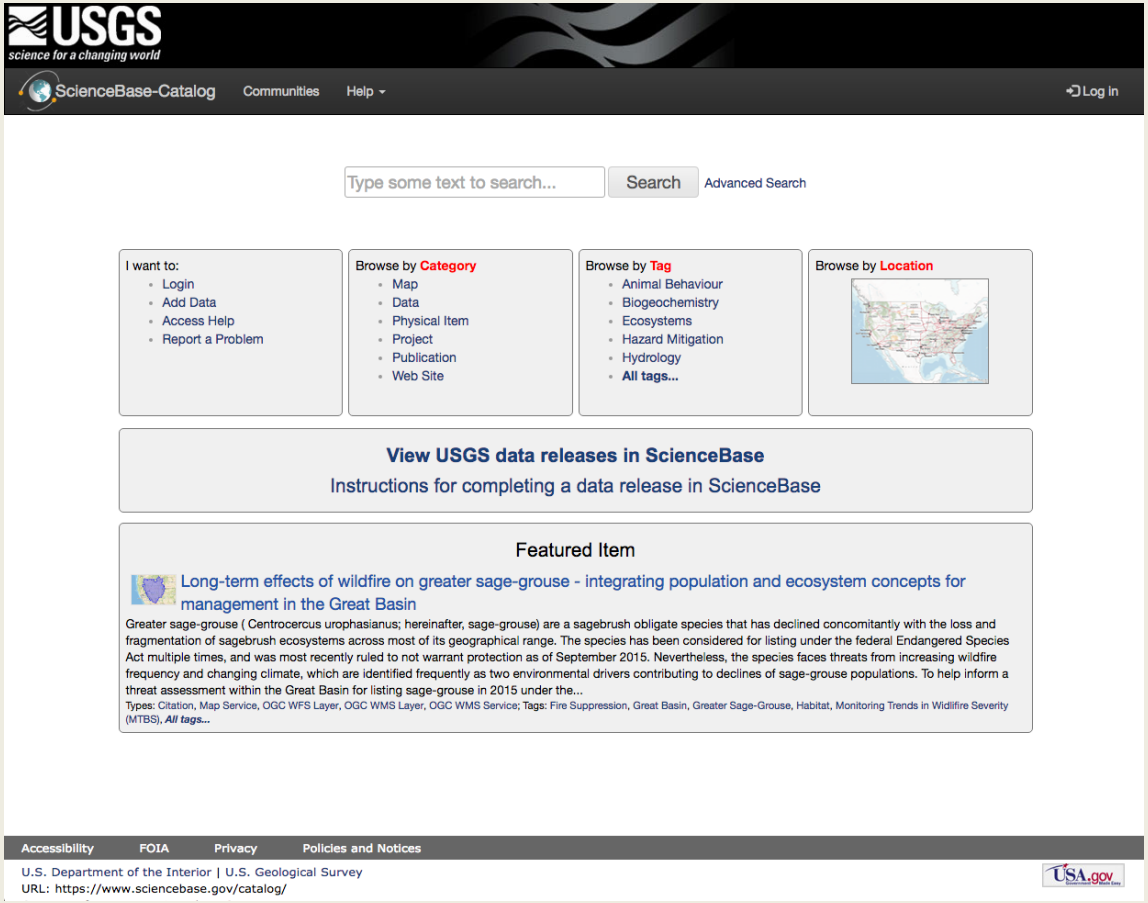
USGS ScienceBase

Science collaboration communities, data and metadata, web services, APIs.

Leveraged for repository, data release platform for USGS researchers

Per Public Access Plan, became a Trusted Digital Repository for USGS

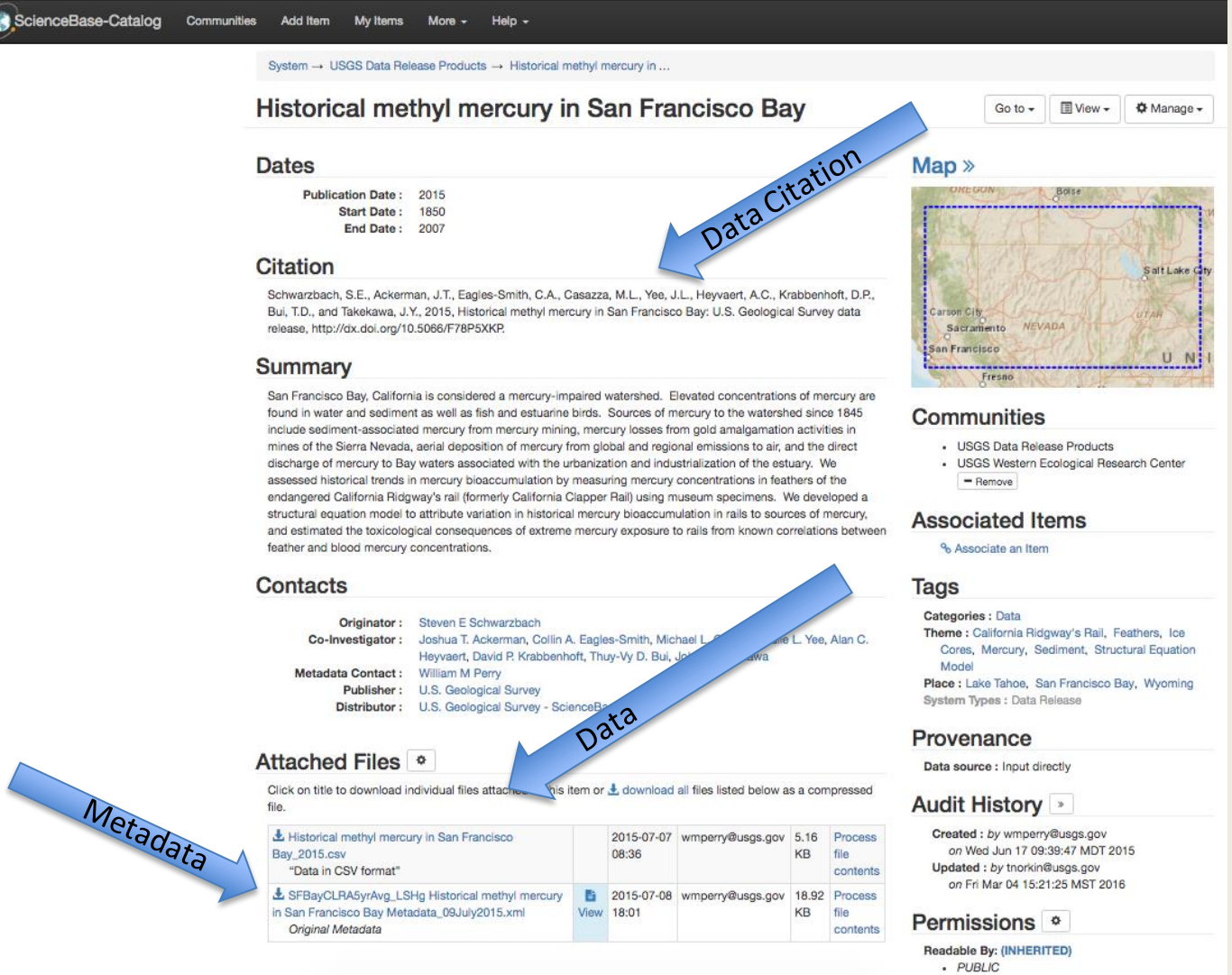
sciencebase.gov



USGS
science for a changing world

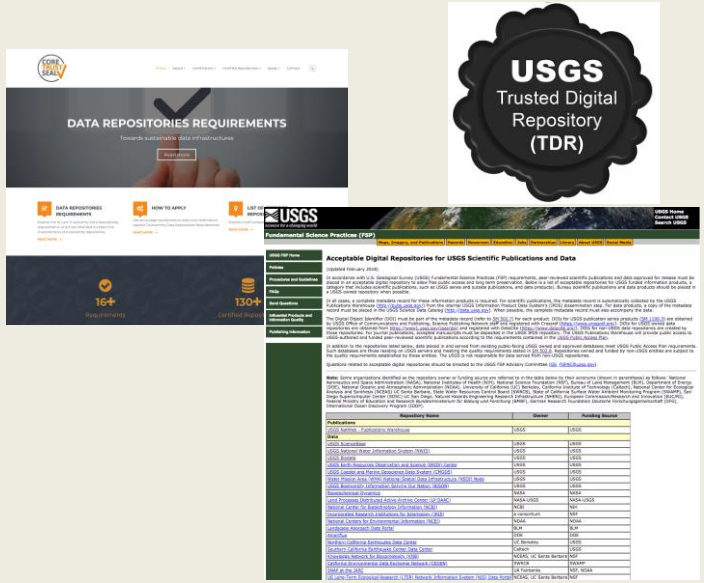
- Data release landing page

  Note the data, metadata, citation, digital object identifier…

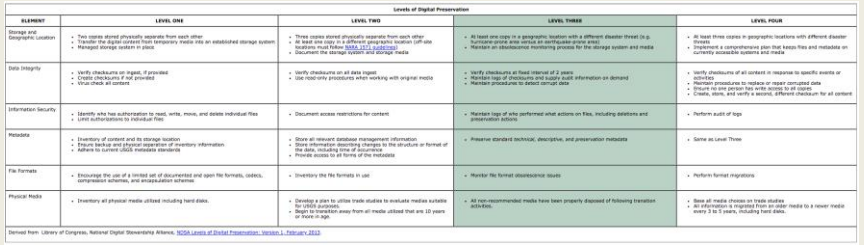


USGS Trusted Digital Repositories

- **USGS is establishing internal USGS TDRs for science data:**
  - Use internationally agreed on criteria from World Data Center Data Seal of Approval-World Data System (now, Core Trust Seal)
  - Review committee approves/disapproves submissions
  - Approvals in place for 3 years - recertification
  - Only an internal process at this time
  - https://www2.usgs.gov/fsp/usgs_trusted_digital_respository.asp



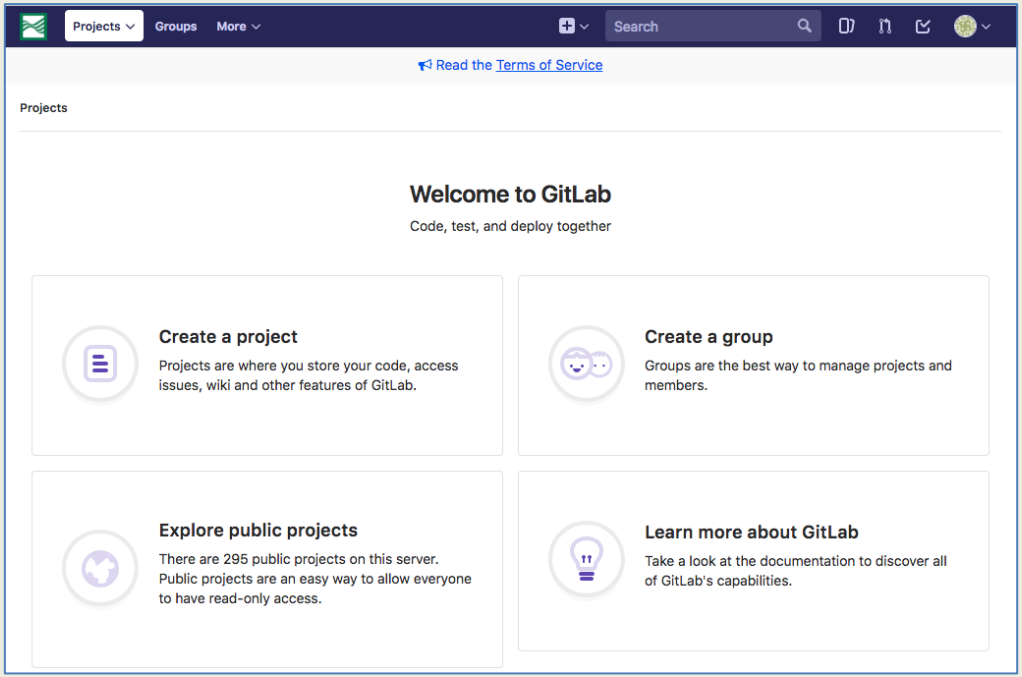
- **Outside repositories: USGS established guidelines for use of repositories outside of USGS**
  - Requires same steps for authors (review and approval, metadata with dois, connection back to SDC, etc)



https://www2.usgs.gov/fsp/trusted_respositories_digital_assets.asp

Scientific Code Release Repository

- Allows code release for USGS researchers
- Per Public Access Plan, software code is required to undergo review and approval under USGS FSP
- Documentation and release of code advances science reproducibility



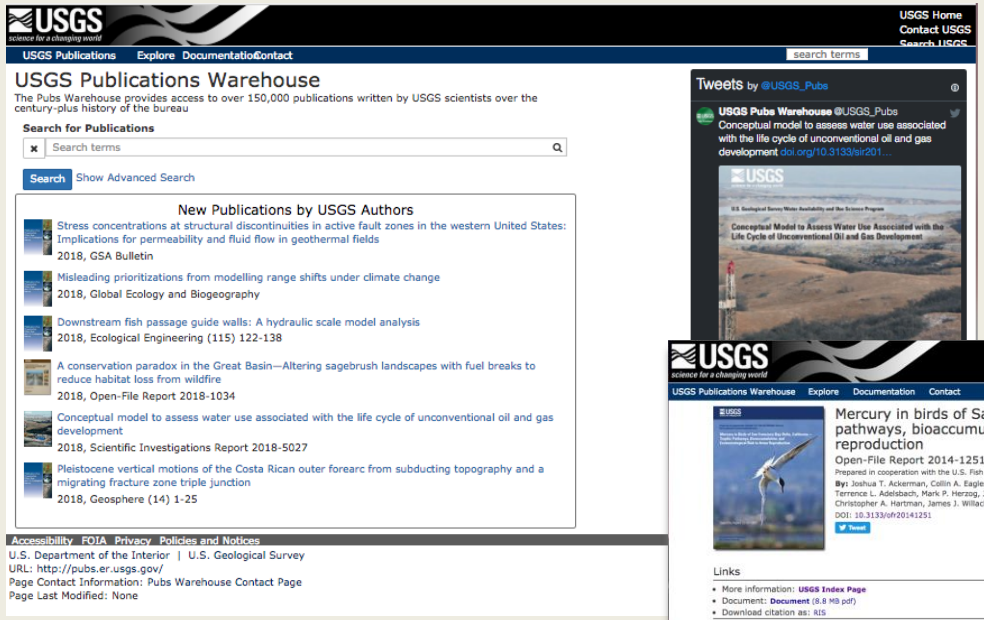
USGS Science Data Catalog

- Catalog of all metadata records describing data released by USGS
- Reporting capability to OMB
- Serves metadata to Department of Interior

data.usgs.gov

USGS Publications Warehouse

- Index to official, USGS-authored publications of the USGS
- Per Public Access Plan, monitors for embargo of publications
- Makes connections to data releases (where applicable)



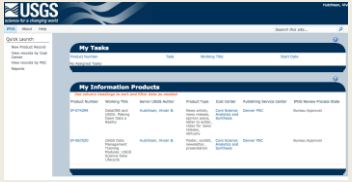
https://pubs.er.usgs.gov/

USGS science for a changing world

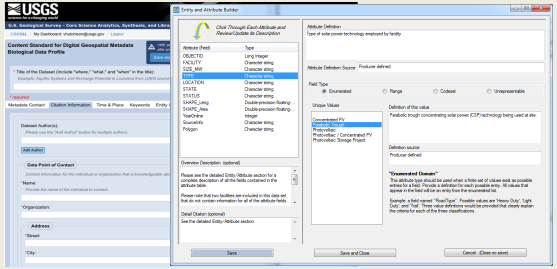
CHORUS

- Clearinghouse for Open Research of the United States
- Per Public Access Plan, USGS uses CHORUS to track public access to publications

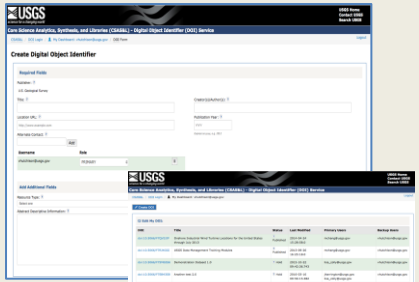
Leveraging and Connecting USGS Tools



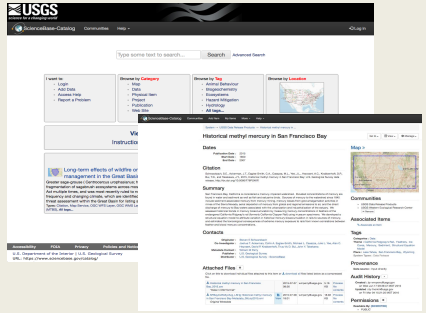
Review/Approval



Metadata



Digital Object Identifiers

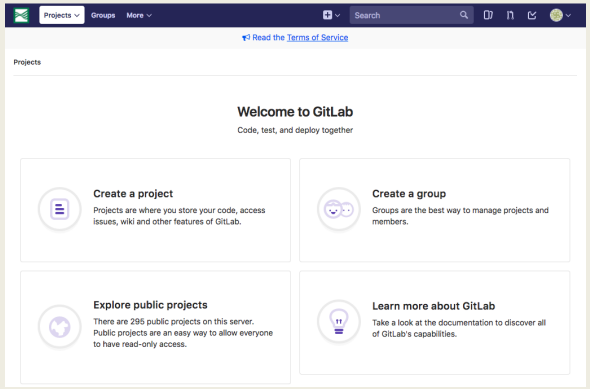


Repositories

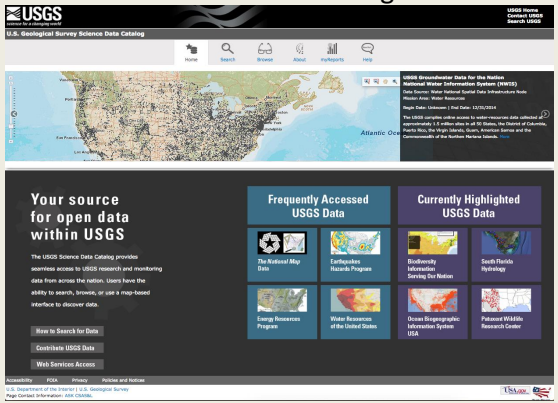
ORCID Connecting Research and Researchers

CHORUS

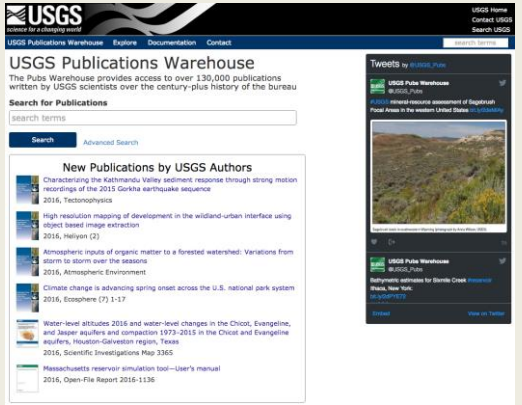
Software Release



Science Data Catalog



Publications Warehouse

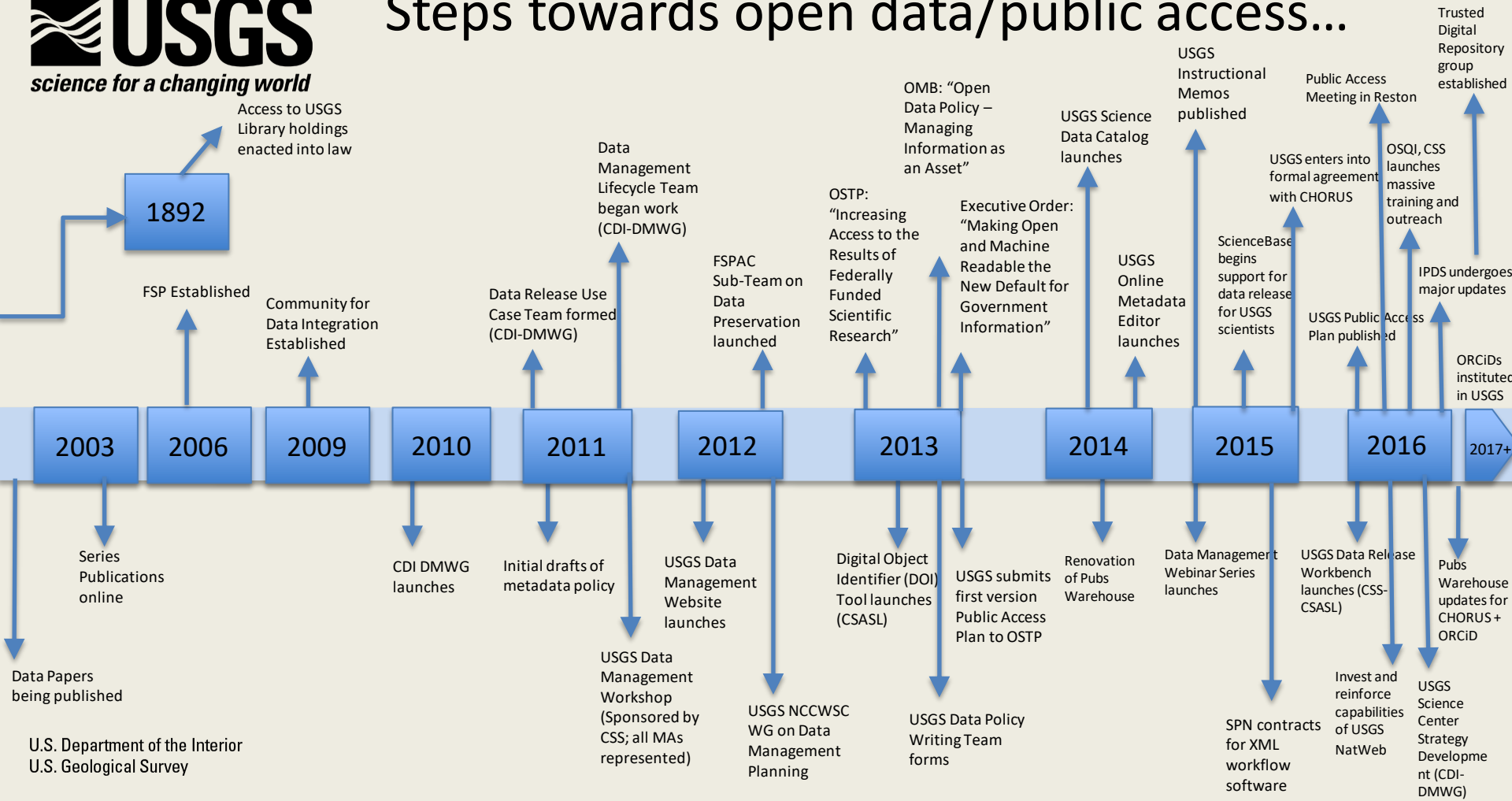


USGS
science for a changing world

USGS Path to Open Science

Communication
Meetings
Trainings

Steps towards open data/public access…



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

Public Access Plan: Communication

- We had 6 months from time of publication of the Public Access Plan to implement
- USGS joined together in a team effort
- Tools application leaders and science programs gathered together in March 2016 to create a ‘road map’ for data release with scholarly publications
- People across the Bureau who had never known each other before began to work together on solutions
- After an initial in-person meeting of ~40 people, we held monthly “accountability” meetings to ensure our work was getting done; coordinators met weekly

We are still working…

Communication and Training

USGS communicated policy and implementation "how-to" together

- Big push to inform Centers/Programs about both policy and data release tools
- Team approach:
  - Office of Science Quality and Integrity (OSQI) – Bureau Approving Officials
  - Core Science Systems (CSS) –Science Analytics Synthesis (SAS) – Science Data Management Team (SDM)

~75

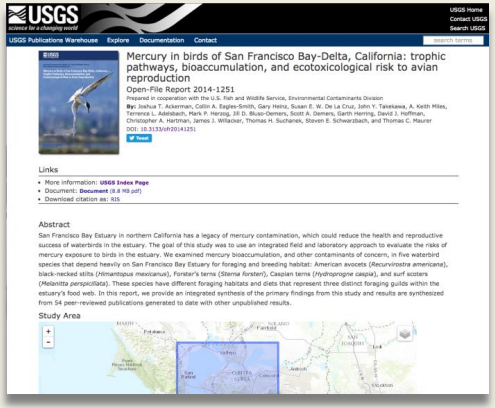
Total presentations between April and June, 2016

~1,610

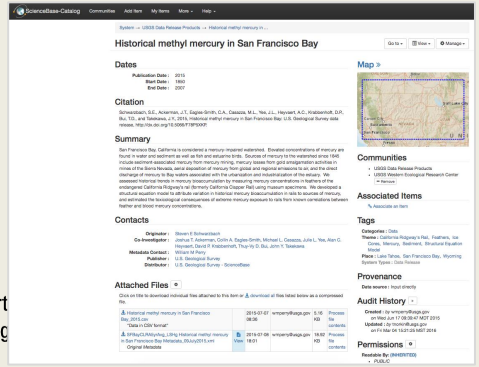
Total # Attendees across all presentations

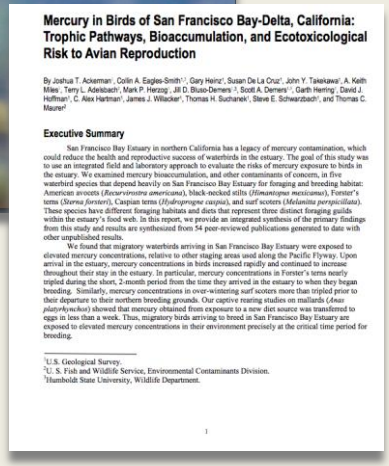
Making Progress: Access to Data & Publications

Publication in PubsWarehouse



Data in ScienceBase



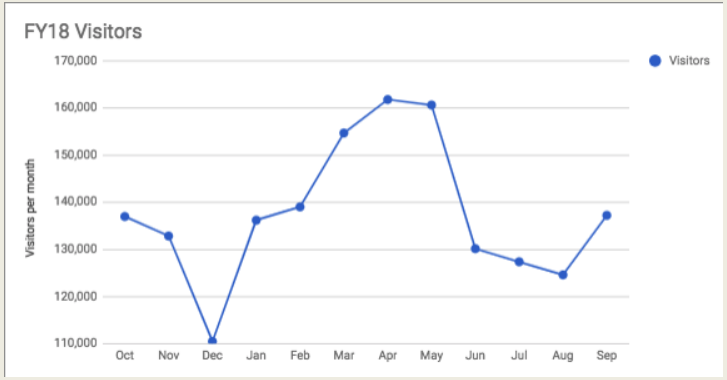


USGS Public Access Plan allows USGS to continue to move forward…

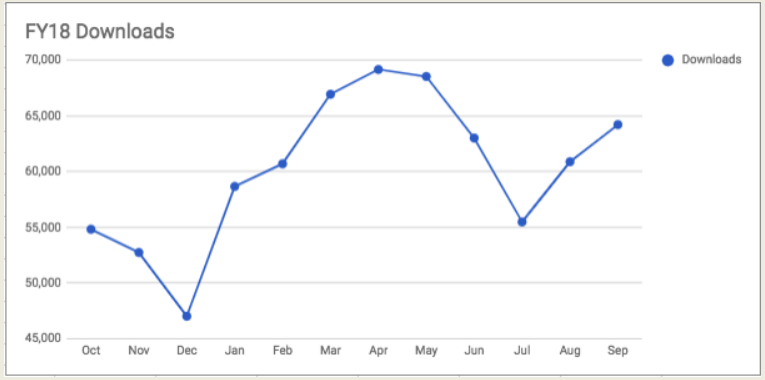
We strive to:

- implement FAIR: Data is findable, accessible, interoperable, and reusable for future research;
- Offer more credit for data release in RGE process
- Formalize data management positions
- Increase data citation rates
- Resolve large data release and preservation challenges

Making Progress: Data and Publications

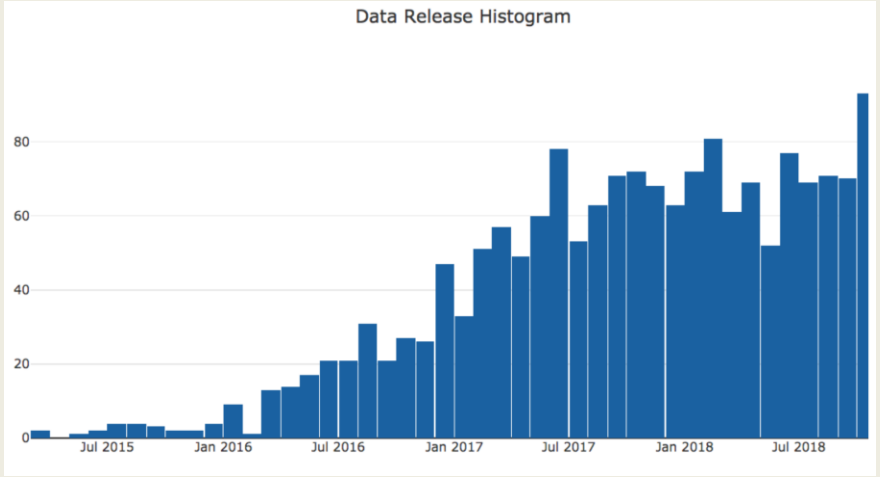


Over 1,651,718 visitors to USGS Publications Warehouse in FY18



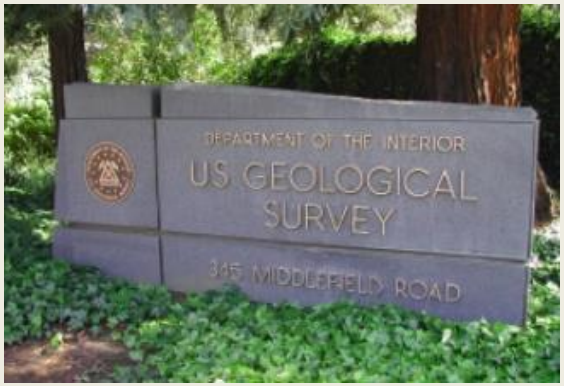
Over 722,332 publication downloads

Over 1,500 USGS data releases are now publicly available in ScienceBase, representing every USGS Mission Area, and 82 different USGS Science Centers and Programs.



USGS
science for a changing world

Thank you!



Questions? Comments?

Viv Hutchison
USGS
vhutchison@usgs.gov