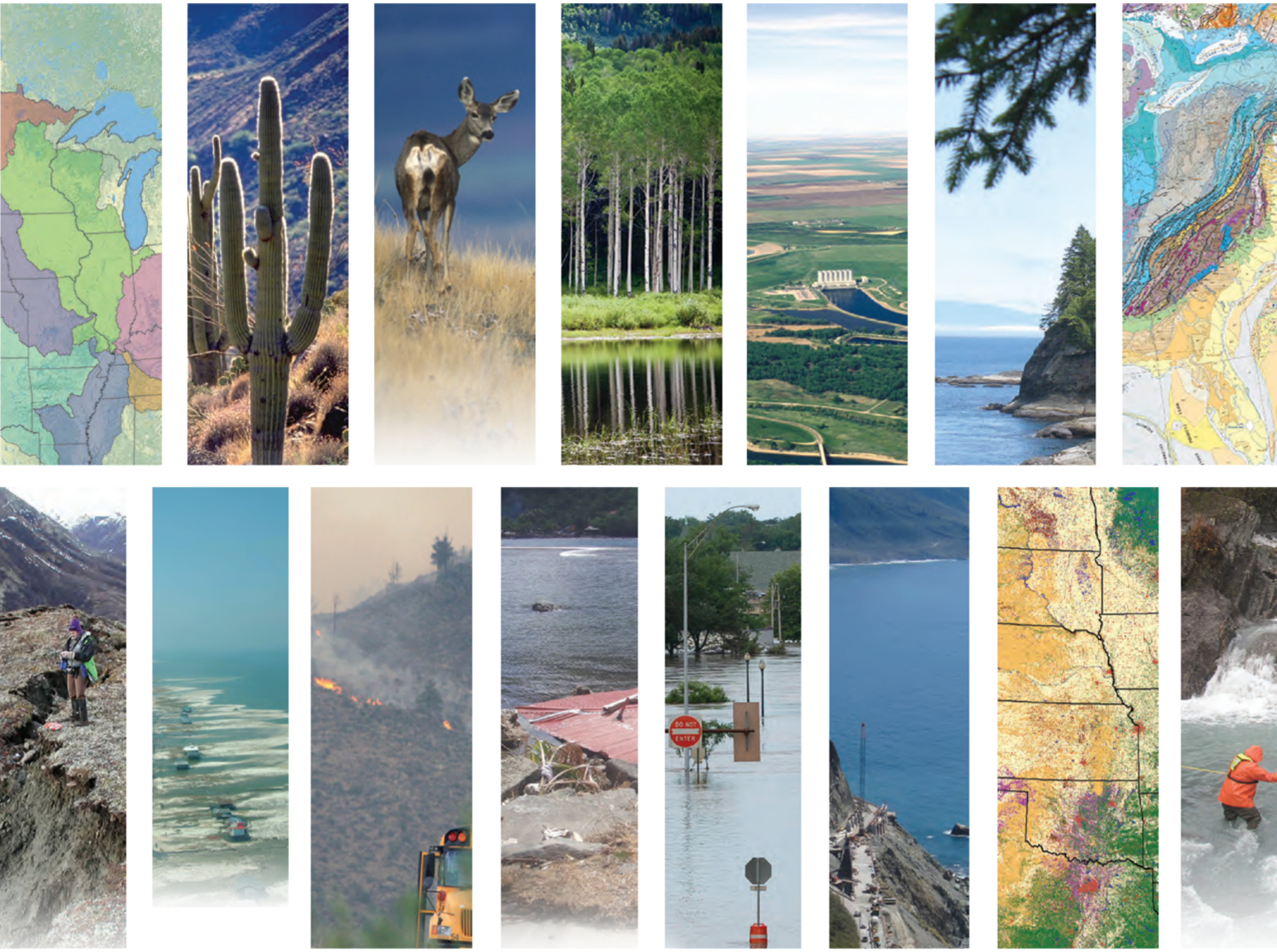
USGS
science for a changing world

Community for Data Integration 2018 Annual Report



Open-File Report 2019–1123

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

Cover. Clockwise from top left. A slice of a map of Mississippi River Basin watersheds derived from hydrologic data available from The National Map (http://www.nationalmap.gov). Cactus in Aqua Fria National Monument, Arizona, December 2012. (Photograph by Bureau of Land Management) Mule deer on rangeland in National Bison Range, Montana, October 2013. (Photograph by Ryan Hagerty, USGS) Aspens reflected in Willow Lake, Wasatch-Cache National Forest, Utah, July 2009. (Photograph by Gary Brewer, USGS) Oahe Dam, South Dakota, 2008. (Photograph by U.S. Army Corps of Engineers) Neah Bay at Cape Flattery, Washington, July 2009. (Photograph by G. Lynn Wingard, USGS) A slice of the Geologic Map of North America available from https://www.usgs.gov/media/images/geologic-map-north-america. Solomon Gulch at top of falls near Valdez, Alaska, October 2010. Minor flood stage from Solomon Lake spill. (Photograph by Jeffery M. Wiles, USGS) A slice of the Gap Analysis Program National Land Cover map, 2009, available from http://gapanalysis.usgs.gov/. Landslides threaten human activities and a marine sanctuary along the coastline at Big Sur, California. The USGS works with the California Department of Transportation to understand landslide issues along this part of the California coast. (Photograph by Mark Reid, USGS) June 2008 flood at Cedar Rapids, Iowa. USGS streamgages and onsite observations by USGS field staff provide critical data to flood forecasts made by NOAA. (Photograph by Don Becker, USGS) A roof washed into the ocean near Fagasa, American Samoa, during the tsunami that hit the island on September 29, 2009. USGS researchers work to increase the understanding of tsunami processes and provide the National Oceanic and Atmospheric Administration with real-time earthquake information to create tsunami warnings. (Photograph by Bruce Jaffe, USGS) Squirrel Creek fire near Laramie, Wyoming. (Photograph by Laura Hallberg, USGS) Hurricane Katrina damaged and destroyed houses on Dauphin Island, Alabama, August 31, 2005. The USGS assesses coastal hazard vulnerability, including the severe beach erosion shown here. (Photograph by USGS) The Denali Fault, Alaska, displaced the ground surface by more than 6 meters on November 4, 2002, in a M 7.9 earthquake. The USGS monitors and studies earthquakes in the United States and around the world, providing warnings and other information to support mitigation and response. (Photograph by Peter Haeussler, USGS)

Community for Data Integration 2018 Annual Report

By Leslie Hsu and Leah Colasuonno

Open-File Report 2019–1123

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

U.S. Department of the Interior
DAVID BERNHARDT, Secretary

U.S. Geological Survey
James F. Reilly II, Director

U.S. Geological Survey, Reston, Virginia: 2019

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

For an overview of USGS information products, including maps, imagery, and publications, visit https://store.usgs.gov.

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Suggested citation:
Hsu, L., and Colasuonno, L., 2019, Community for Data Integration 2018 annual report: U.S. Geological Survey Open-File Report 2019–1123, 26 p., https://doi.org/10.3133/ofr20191123.

ISSN 2331-1258 (online)

Contents

Figures

Tables

Abbreviations

3DEP	3D Elevation Program
CDI	Community for Data Integration
DevOps	Software Development and Information Technology Operations
EDGE	Equipment and Development Grade Evaluation
eDNA	Environmental DNA
FY	fiscal year
GIS	Geographic Information Systems
ICEMM	Interagency Collaborative for Environmental Modeling and Monitoring
lidar	light detection and ranging
NAS	Nonindigenous Aquatic Species
USGS	U.S. Geological Survey

Community for Data Integration 2018 Annual Report

By Leslie Hsu and Leah Colasuonno

Abstract

The Community for Data Integration (CDI) is a community of practice whose purpose is to build the U.S. Geological Survey knowledge base in data integration. This annual report describes the various presentations, activities, and outcomes of the CDI monthly forums, working groups, trainings, and other CDI-sponsored events in fiscal year 2018. The report also describes the objectives of the 10 CDI-funded projects for the year. The CDI had a topical theme for fiscal year 2018—Risk assessment and hazards vulnerability in support of integrated predictive science capacity. This report describes how the community coordinated its activities around this theme.

Introduction

The Community for Data Integration (CDI) is a community of practice whose purpose is to build the U.S. Geological Survey (USGS) knowledge base in data integration for the Earth and biological sciences. The CDI accomplishes this by creating an environment to help members increase their expertise in working with scientific data. The CDI focuses on opportunities to share information across disciplines and organizational structures, invigorating cross-boundary communication. Through these efforts, community members are informed of emerging technologies and topics that help them in their professional responsibilities.

The CDI is funded and led by the USGS, but membership is voluntary and open to USGS employees and other individuals and organizations willing to contribute to the community. In fiscal year (FY) 2018, the CDI welcomed 213 new members to its community of learning and sharing, bringing total membership to 1,065. Funding and guidance for the CDI come from the USGS Core Science Systems Mission Area and the USGS Office of Enterprise Information.

The goal of the CDI is to advance the understanding of Earth systems by

- creating and supporting a community of people interested in sharing strategies, methods, tools, and infrastructure, so that members can increase their expertise;
- advocating for practices that support the integration of science information across disciplines and organizational structures;
- supporting innovative ideas through seed-funded projects; and
- developing and providing training opportunities that support data and science integration activities.

A recent focus of the CDI is building what the USGS calls Integrated Predictive Science Capacity (fig. 1). This focus was introduced at the May 2017 CDI workshop (Hsu and others, 2018b). CDI sponsor and Associate Director for Core Science Systems Kevin T. Gallagher noted that the societally significant issues that the USGS addresses are complex and solutions need to be interdisciplinary in order to deliver useful products and information for decision making (Hsu and others, 2018b). By integrating science and information from different disciplines, the USGS and CDI have the expertise to build high-profile, modular components of an integrated decision-support system.

USGS guidance for FY 2018 included additional details about integrated predictive science capacity (see fig. 1), as follows:

> “In 2018 the USGS will begin efforts to design and pilot an integrated scientific capacity to deliver powerful new products and services that provide: 1) vulnerability detection and assessment, 2) prediction and forecasting, 3) early warning, and 4) decision support at the scale of decisions. Examples of what could be achieved include seasonal to real-time warnings of: biological threats such as disease, invasive species, or harmful algal blooms; natural hazards, such as earthquakes, landslides, volcanic eruptions, and flood inundation; impacts of both sudden and long-term coastal change on public safety, infrastructure and economies, and lands, waters and natural resources; health threats from environmental contaminants and pathogens; and water availability or quality prediction and forecasting.

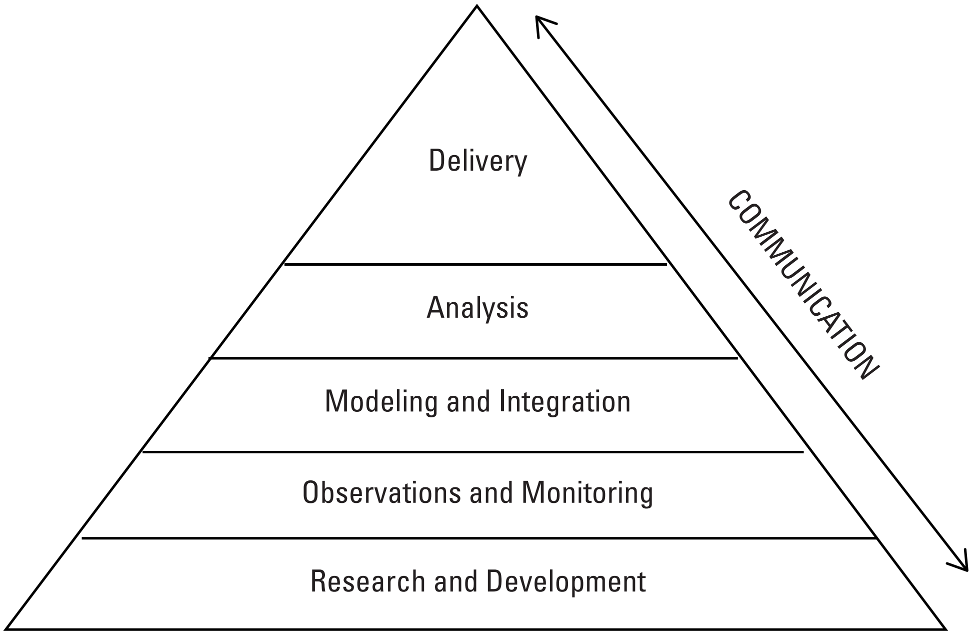


Figure 1. Integrated Predictive Science Capacity triangle.

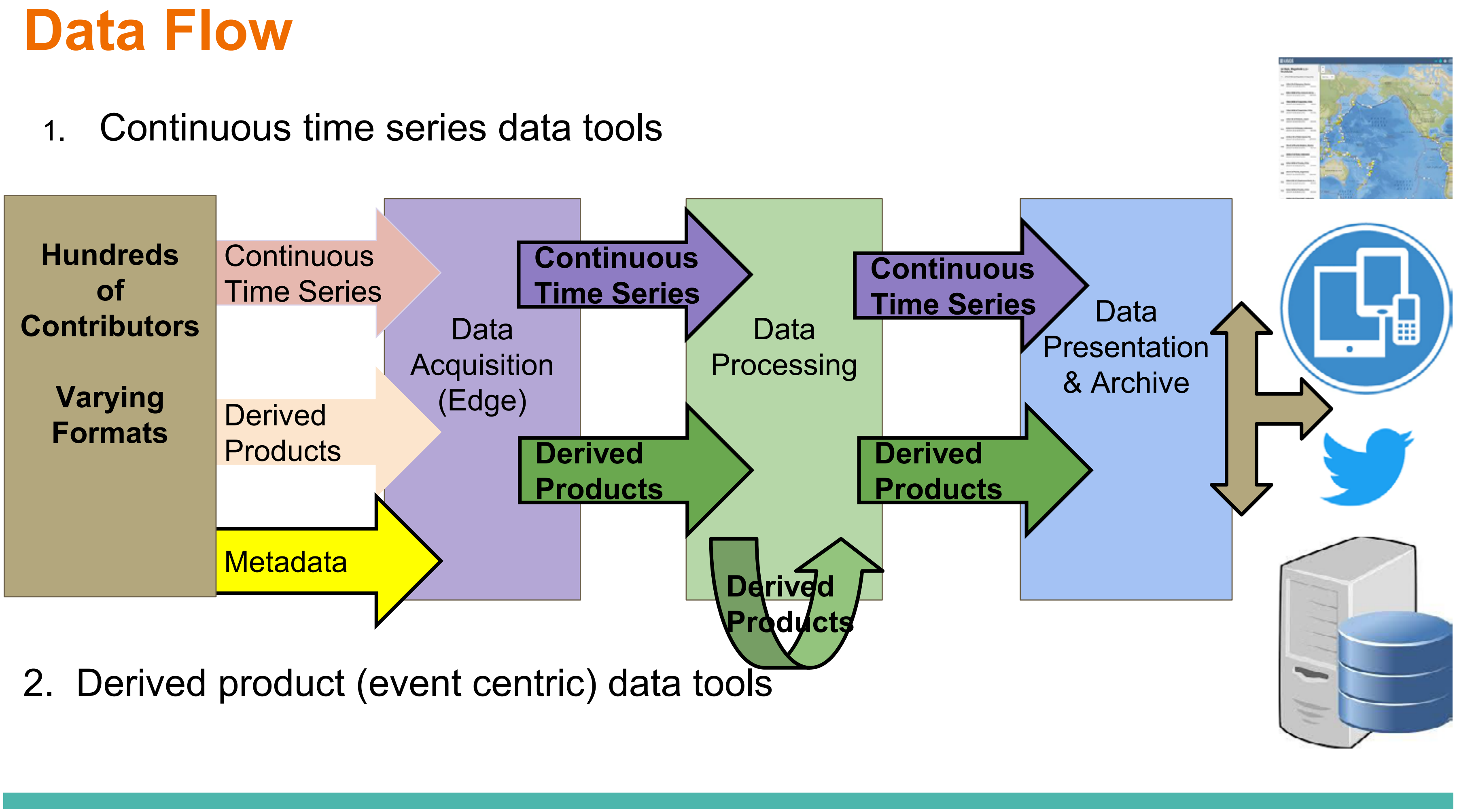
> "This integrated capacity will span scientific boundaries and disciplines, and require investments in data integration, high performance computing, modeling, analytics, laboratory facilities, and visualization and decision support tools. This integration of science, facilities, data, models, and tools will provide enhanced and tangible value to the Nation, secure USGS leadership in earth and natural science, and simultaneously create the building blocks for a more fully-integrated science agency" (USGS, 2017b).

In FY 2018, the CDI focused on this theme of integrated predictive science capacity and followed with the pilot theme of risk assessment and hazard vulnerability. This was accomplished in several ways: by focusing on these two themes in the annual CDI proposals process; by funding a team that leveraged an existing U.S. Department of the Interior project (see Wood and others, 2019) to develop a CDI risk map and underlying tools for researchers; by having a special workshop in January 2018 to bring together USGS stakeholders to discuss risk assessment and hazard vulnerability; and by coordinating meetings and presentations of risk-related projects, including at external meetings such as the Earth Science Information Partners Summer Meeting. The following sections lay out more detail of the events, outputs, and impact of the CDI in FY 2018.

Monthly Forums

The CDI's monthly virtual forums enable community members to stay informed of new tools, best practices, standards, and policies within the Earth and biological sciences community. Topics in FY 2018 included the Freedom of Information Act at the USGS, practices at USGS computing centers such as the National Earthquake Information Center (fig. 2), new mobile applications that inform users of flood conditions, services for USGS researchers in cloud computing and large data management, and data visualization platforms (fig. 3). Many of the presentations reported on the outputs of previously funded CDI projects from FY 2017. Table 1.1 in the appendix lists the presentations and speakers. Participation in monthly meetings increased by 20 percent from the previous year to an average of 97 online participants.

One focus in FY 2018 was the reproducible notebook series, which was proposed at the 2017 CDI workshop (Hsu and others, 2018b) and led by Rich Signell, research oceanographer at the Woods Hole Coastal and Marine Science Center. The series demonstrates interoperability by accessing and using data available from different web services across disciplines. The presentations demonstrated that we have made some progress on interoperable web services and provided examples that could be reused for other service and interoperability testing. Topics included access of Ocean Biogeographic Information System (OBIS) data, streamlining tasks in data management, using USGS Geo Data Portal services, and publishing Jupyter notebooks (fig. 4).



Data Flow

1. Continuous time series data tools

Hundreds of Contributors

Varying Formats

Continuous Time Series

Derived Products

Metadata

Data Acquisition (Edge)

Continuous Time Series

Derived Products

Data Processing

Derived Products

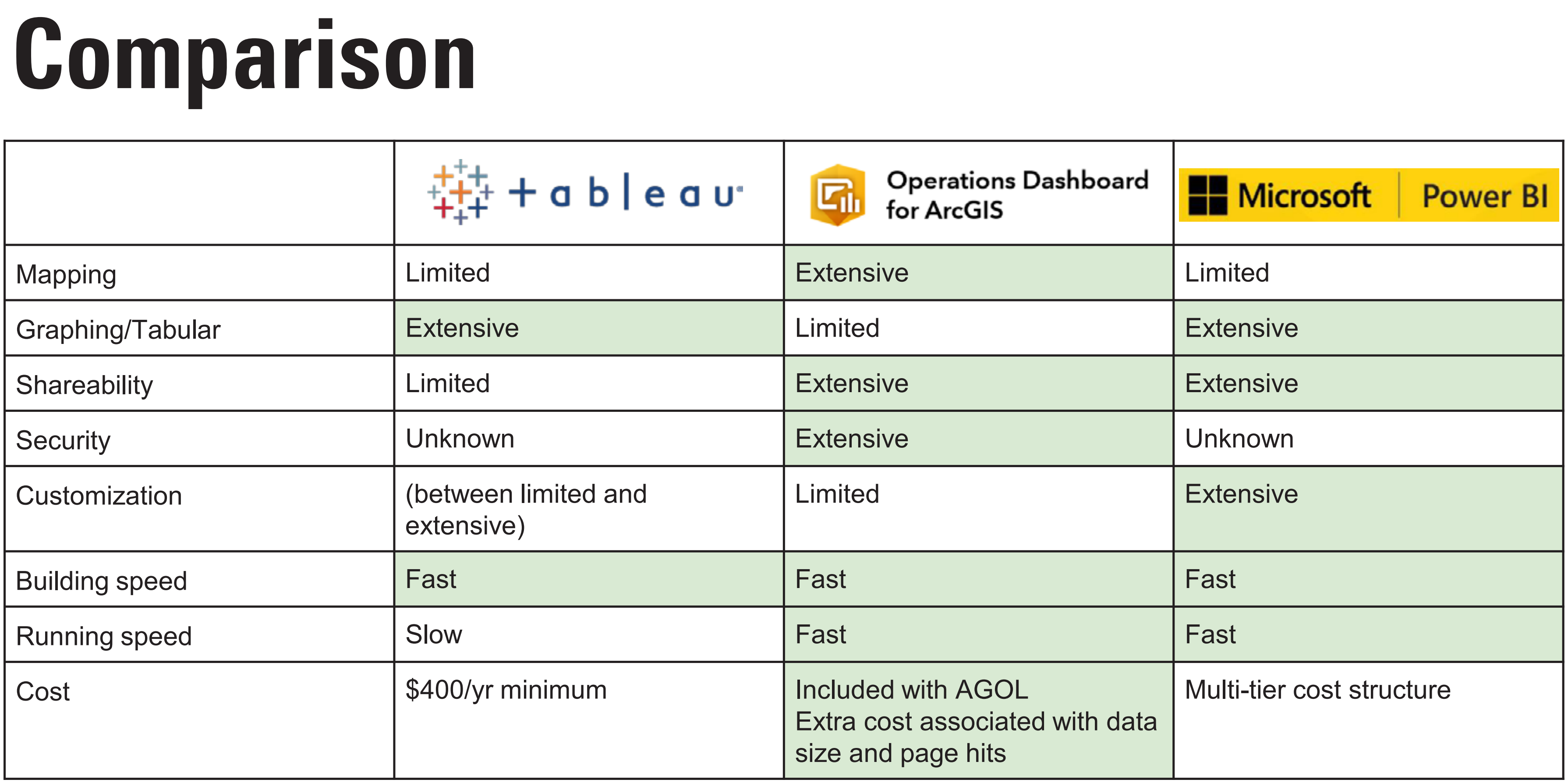
Continuous Time Series

Derived Products

Data Presentation & Archive

2. Derived product (event centric) data tools

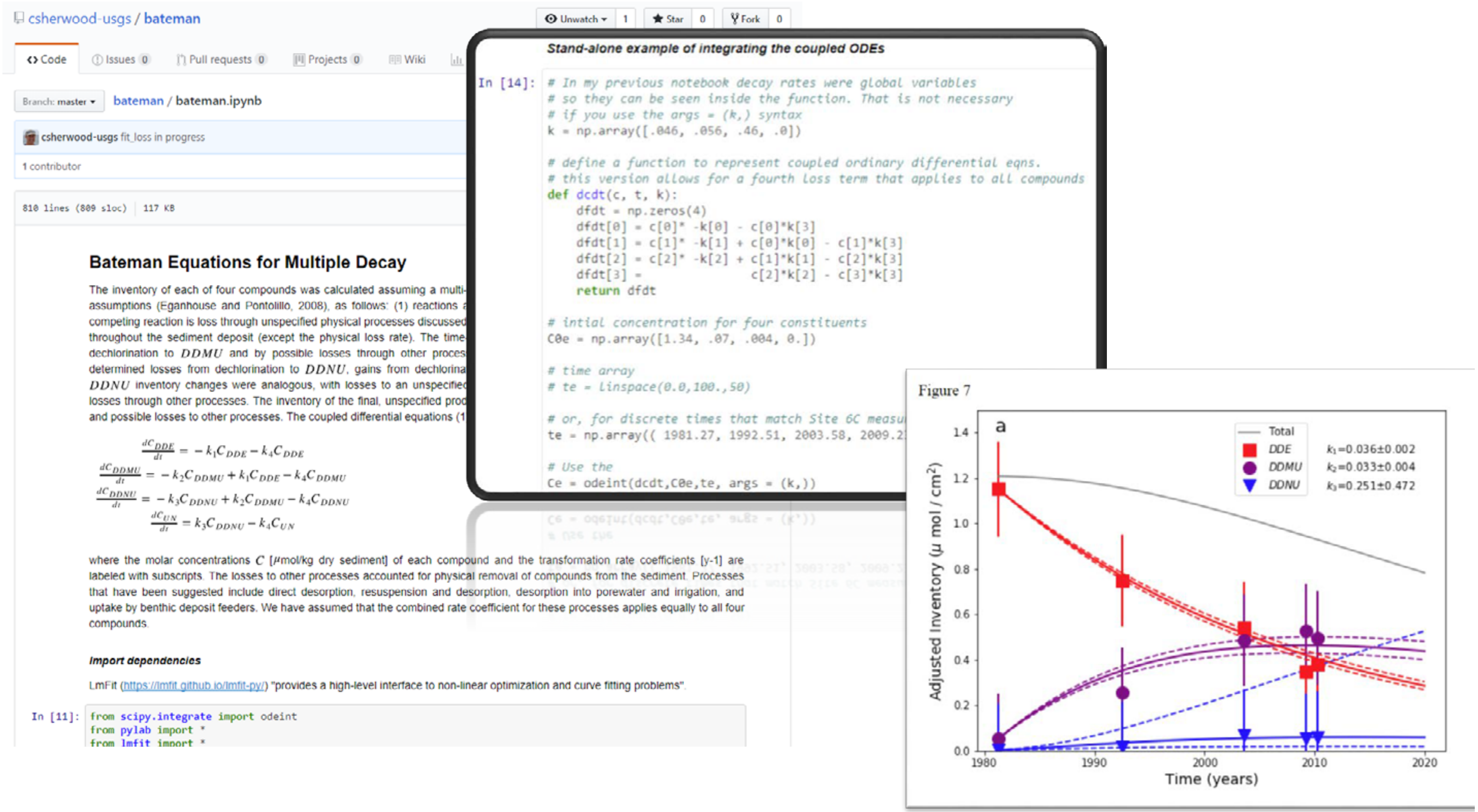
Figure 2. Slide from the November 2017 Monthly Forum presentation “National Earthquake Information Center: Overview of Real-Time Data Acquisition, Processing, and Archive” (Michelle Guy).



Comparison

	tableau	Operations Dashboard for ArcGIS	Microsoft Power BI
Mapping	Limited	Extensive	Limited
Graphing/Tabular	Extensive	Limited	Extensive
Shareability	Limited	Extensive	Extensive
Security	Unknown	Extensive	Unknown
Customization	(between limited and extensive)	Limited	Extensive
Building speed	Fast	Fast	Fast
Running speed	Slow	Fast	Fast
Cost	$400/yr minimum	Included with AGOL Extra cost associated with data size and page hits	Multi-tier cost structure

Figure 3. Slide from the June 2018 presentation “Data Visualization for Science: Comparing 3 Dashboard Building Software Packages” (Kevin Henry, Jason Sherba, and Jeff Peters). (yr, year)



Bateman Equations for Multiple Decay

The inventory of each of four compounds was calculated assuming a multi-
assumptions (Eganhouse and Pontolillo, 2008), as follows: (1) reactions a
competing reaction is loss through unspecified physical processes discussed
throughout the sediment deposit (except the physical loss rate). The time-
dechlorination to $DDMU$ and by possible losses through other proces
determined losses from dechlorination to $DDNU$, gains from dechlorina
$DDNU$ inventory changes were analogous, with losses to an unspecified
losses through other processes. The inventory of the final, unspecified prod
and possible losses to other processes. The coupled differential equations (1

$$\frac{dC_{DDE}}{dt} = -k_1 C_{DDE} - k_4 C_{DDE}$$

$$\frac{dC_{DDMU}}{dt} = -k_2 C_{DDMU} + k_1 C_{DDE} - k_4 C_{DDMU}$$

$$\frac{dC_{DDNU}}{dt} = -k_3 C_{DDNU} + k_2 C_{DDMU} - k_4 C_{DDNU}$$

$$\frac{dC_{UN}}{dt} = k_3 C_{DDNU} - k_4 C_{UN}$$

where the molar concentrations C [μmol/kg dry sediment] of each compound and the transformation rate coefficients [y-1] are labeled with subscripts. The losses to other processes accounted for physical removal of compounds from the sediment. Processes that have been suggested include direct desorption, resuspension and desorption, desorption into porewater and irrigation, and uptake by benthic deposit feeders. We have assumed that the combined rate coefficient for these processes applies equally to all four compounds.

Import dependencies

LmFit (https://lmfit.github.io/lmfit-py/) "provides a high-level interface to non-linear optimization and curve fitting problems".

```
In [11]: from scipy.integrate import odeint
         from pylab import *
         from lmfit import *
```

Stand-alone example of integrating the coupled ODEs

```
In [14]: # In my previous notebook decay rates were global variables
         # so they can be seen inside the function. That is not necessary
         # if you use the args = (k,) syntax
         k = np.array([.046, .056, .46, .0])

         # define a function to represent coupled ordinary differential eqns.
         # this version allows for a fourth loss term that applies to all compounds
         def dcdt(c, t, k):
             dfdt = np.zeros(4)
             dfdt[0] = c[0]* -k[0] - c[0]*k[3]
             dfdt[1] = c[1]* -k[1] + c[0]*k[0] - c[1]*k[3]
             dfdt[2] = c[2]* -k[2] + c[1]*k[1] - c[2]*k[3]
             dfdt[3] =                c[2]*k[2] - c[3]*k[3]
             return dfdt

         # intial concentration for four constituents
         C0e = np.array([1.34, .07, .004, 0.])

         # time array
         # te = linspace(0.0,100.,50)

         # or, for discrete times that match Site 6C measu
         te = np.array(( 1981.27, 1992.51, 2003.58, 2009.2

         # Use the
         Ce = odeint(dcdt,C0e,te, args = (k,))
```

Figure 4. Slide from the January 2018 Reproducible Notebook Series presentation "Publishing Jupyter Notebooks" (Christopher Sherwood).

Collaboration Areas

The CDI is organized into groups, or collaboration areas, that form around common interests in specific topics related to data integration (table 1). These groups provide a platform for sharing resources and knowledge, discussing challenges, and identifying solutions that will help to advance data integration in the Earth and biological sciences. Collaboration area membership is voluntary and open to anyone interested in participating. In FY 2018, five collaboration areas began their activities within the CDI: Citizen-Centered Innovation, Environmental DNA (eDNA), Elevation, Geographic Information Systems (GIS), and Subduction Zone Science. In addition, work continued in FY 2018 in the Bioinformatics, Data Management, Data Science, DevOps (Software Development and Information Technology Operations), Earth-Science Themes, Interagency Collaborative for Environmental Modeling and Monitoring (ICEMM), Metadata Reviewers, Semantic Web, Software Development, and Technology Stack collaboration areas.

"Collaboration area" is an over-arching term that includes different names the groups have chosen for themselves. "Community of practice," "working group," "focus group," and "cluster" are different names that reflect the differing group goals as well as the naming convention at the time of group formation (See Glossary for definitions). Collaboration area discussions may originate in their separate groups but are often widely applicable to a larger audience. Some recent examples include the discussion of Federal Source Code policy for scientific software, licensing of software, metadata and metadata reviews and tools, tools and new data formats for computing in the cloud, and the EDGE (Equipment and Development Grade Evaluation) development path. Collaboration area activities are documented in the wiki for use by a wider audience as much as possible (https://my.usgs.gov/confluence/x/yhv1I). A brief description of each collaboration area and its activities in FY 2018 is provided in the following sections.

Table 1. Community for Data Integration collaboration areas with activity in fiscal year 2018 and contacts.

[DevOps, Software Development and Information Technology Operations]

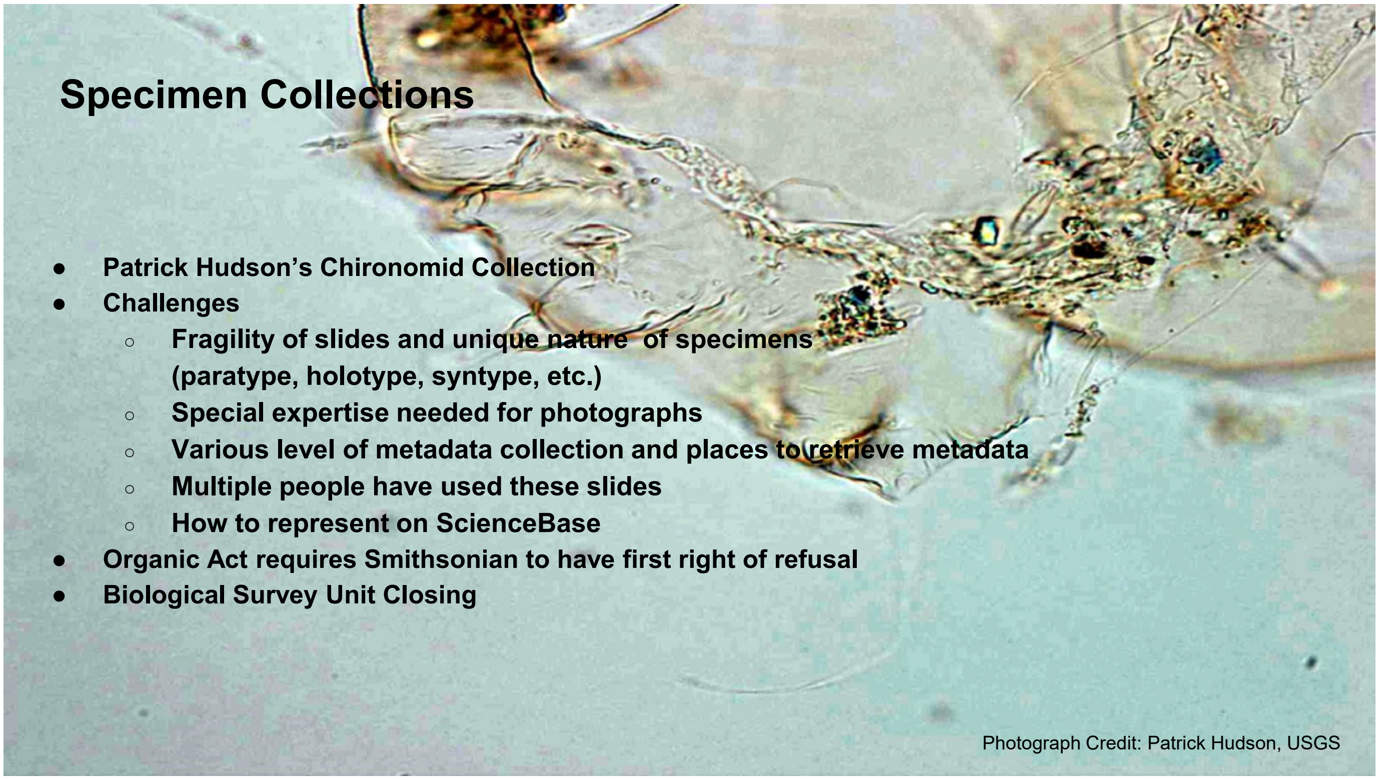
Collaboration area topic	Group contact(s)
Citizen-Centered Innovation	Sophia Liu—sophialiu@usgs.gov
Data Management	Viv Hutchison—vhutchison@usgs.gov Cassandra Ladino—ccladino@usgs.gov
Data Science	Lindsay Carr—lcarr@usgs.gov
DevOps	David Hughes—drhughes@usgs.gov
Earth-Science Themes	Roland Viger—rviger@usgs.gov
Bioinformatics	Scott Cornman—rcornman@usgs.gov Christina Kellogg—ckellogg@usgs.gov Denise Akob—dakob@usgs.gov
eDNA	Pete Ruhl—pmruhl@usgs.gov JC Nelson—jcnelson@usgs.gov
Elevation	Cynthia Miller-Corbett—cmcorbet@usgs.gov JC Nelson—jcnelson@usgs.gov Jason Stoker—jstoker@usgs.gov
Subduction Zone	Joan Gomberg—gomberg@usgs.gov
Geographic Information Systems	Shane Wright—wright@usgs.gov Andy Lamotte—alamotte@usgs.gov Roland Viger—rviger@usgs.gov
Interagency Collaborative for Environmental Modeling and Monitoring	Pierre Glynn—pglynn@epa.gov
Metadata Reviewers	Fran Lightsom—flightsom@usgs.gov
Semantic Web	Fran Lightsom—flightsom@usgs.gov
Software Development	Blake Draper—bdraper@usgs.gov Michelle Guy—mguy@usgs.gov
Metadata Reviewers	Fran Lightsom—flightsom@usgs.gov
Technology Stack	Richard Signell—rsignell@usgs.gov

Citizen-Centered Innovation

The Citizen-Centered Innovation Community provides information to its members on open innovation efforts like crowdsourcing, citizen science, civic hacking, and challenge and prize competitions at USGS and other U.S. Department of the Interior bureaus. The group facilitates and enhances connections between the USGS and the larger Federal and public Citizen Science and Open Innovation communities. At their meetings, they share best practices, resources, and emerging open innovation projects (appendix table 1.2). The group provided information to the CitizenScience.gov website about USGS crowdsourcing and citizen science projects (https://www.citizenscience.gov/catalog/usgs/#) and collected input to documents such as the "Report to Congress for the Crowdsourcing and Citizen Science Act (15 U.S.C. 3724)" (Office of Science & Technology Policy, 2019).

Data Management

The Data Management Working Group fosters best practices and collaborative approaches for incorporating data management into USGS science and educating scientists about the value of data management. The group seeks to elevate the practice of data management such that it is seen as a critical part of the pursuit of science in the USGS. In FY 2018, the Data Management Working Group continued to keep its members up to date on new tools and methods as well as data policy and procedures in the USGS. The group hosted a series of presentations on 21 topics including "Data Management Activities in the Water Mission Area;" "Data Integration, Fiscal Accountability, and the 'Business of Science;'" "Taking Action Against USGS Legacy Data Challenges" (fig. 5); "Capturing Your Processing and Analysis Workflow in R" (fig. 6); and "Tidy Data—Using Python and Pandas to Clean Legacy Datasets" (appendix table 1.3).

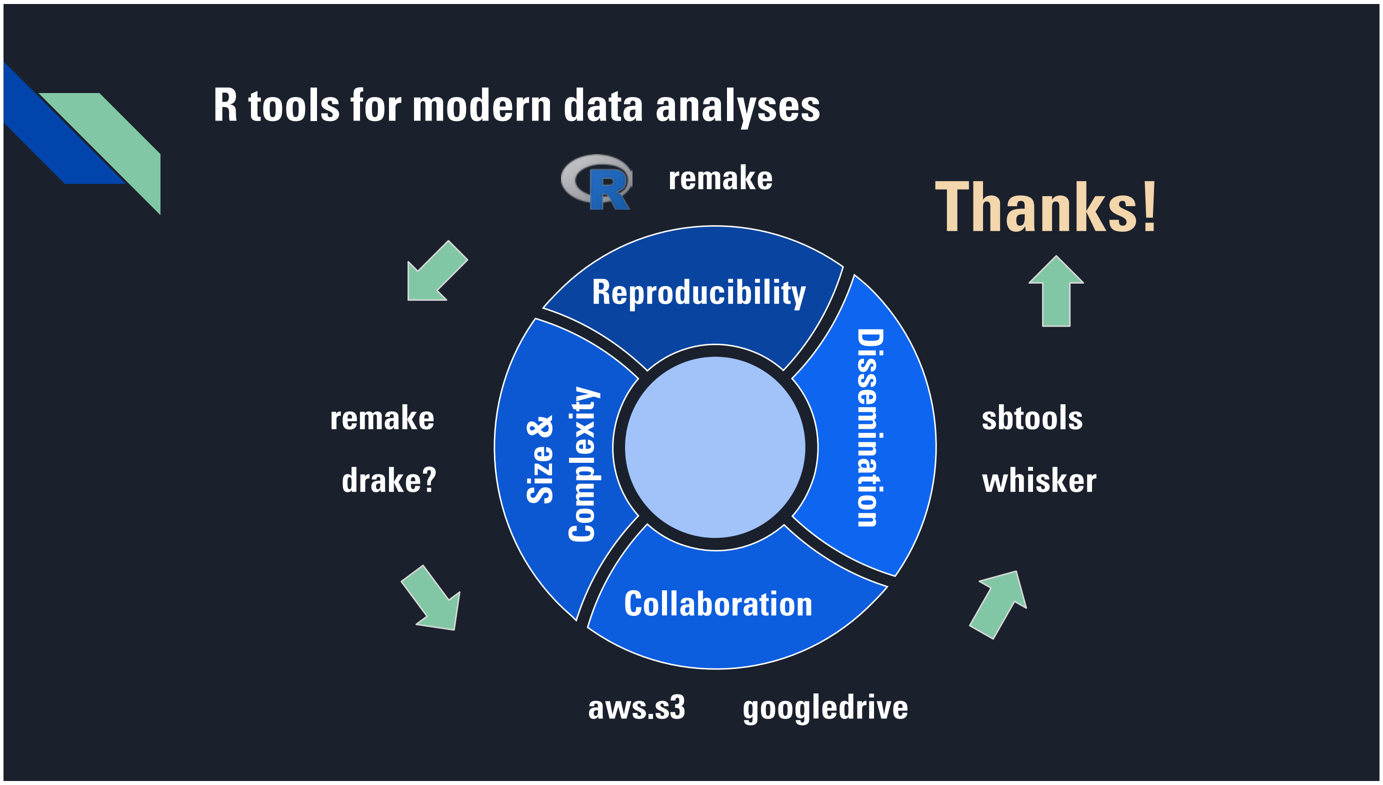


Specimen Collections

- Patrick Hudson's Chironomid Collection
- Challenges
  - Fragility of slides and unique nature of specimens (paratype, holotype, syntype, etc.)
  - Special expertise needed for photographs
  - Various level of metadata collection and places to retrieve metadata
  - Multiple people have used these slides
  - How to represent on ScienceBase
- Organic Act requires Smithsonian to have first right of refusal
- Biological Survey Unit Closing

Photograph Credit: Patrick Hudson, USGS

Figure 5. Slide from the April 2018 presentation "Taking Action Against USGS Legacy Data Challenges" (Tara Bell and Lance Everette). (USGS, U.S. Geological Survey)



R tools for modern data analyses

remake

Thanks!

Reproducibility

Size & Complexity

Dissemination

Collaboration

remake

drake?

sbtools

whisker

aws.s3

googledrive

Figure 6. Slide from the March 2018 presentation “Capturing Your Processing and Analysis Workflow in R” (Alison Appling).

Data Science

The purpose of the Data Science Community of Practice is to share content related to data science at the USGS. For purposes of the CDI, data science is defined as the application of computer science, machine learning, data visualization, and other emerging technical approaches to enhance more traditional USGS science. The group maintains a directory for data science enthusiasts on the CDI wiki space. The Data Science Community of Practice does not have regular meetings. Instead, they communicate through forums on GitHub (https://github.com/usgs/best-practices) and on USGS Slack (#data-science).

DevOps

The purpose of the DevOps Working Group is to share new techniques and lessons learned using DevOps tools and methods. DevOps stands is an abbreviation for Software Development and Information Technology Operations and the group aims to improve efficiency by unifying software development and software operation, which have traditionally been separate tasks in organizations. The group facilitates communication across organizational, regional, and managerial boundaries, allowing USGS project managers and information technology and development staff to share how DevOps-related methods, techniques, and tools are enabling their local activity. In FY 2018, the group hosted presentations on topics such as USGS Git options, automating Esri services with Jenkins, California Polytechnic State University Amazon Web Services migration and activities, experiences with Agile contracts, and a software inventory for USGS (appendix tables 1.4a and 1.4b).

Earth-Science Themes

The goal of the Earth-Science Themes Working Group is to provide a forum for applied Earth science within the CDI. An additional goal of the group is to bring fundamental Earth science data producers, such as the USGS National Hydrography Dataset, 3D Elevation Program (3DEP), and Multi-Resolution Land Characteristics Program, into more direct and regular contact with scientists who work to integrate the sometimes independent data sources developed by these programs. In FY 2018, the Earth-Science Themes Working Group provided an umbrella for the distinctive themes of Bioinformatics, eDNA, Elevation, Water, Risk, and Subduction Zone Science.

Bioinformatics

The Bioinformatics Community of Practice meets to discuss bioinformatic tools, methods, resources, and data handling techniques. Topics in FY 2018 included efforts to provide templates and guidance for genetic data release, an overview of microbiome studies at the USGS (fig. 7; Kellogg and Hopkins, 2017), a demonstration of the iMicrobe research platform, and methods for analyzing single nucleotide polymorphism (SNP) datasets (appendix table 1.5).

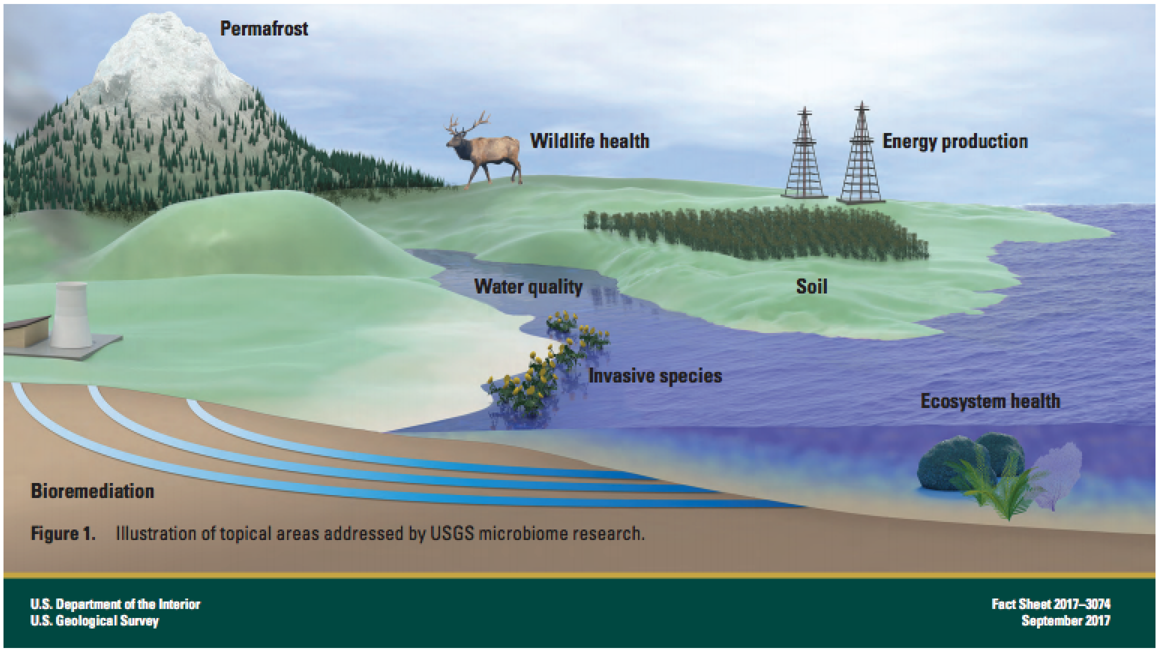


Figure 7. Image from a U.S. Geological Survey (USGS) fact sheet on microbiome research (Kellogg and Hopkins, 2017) that was discussed at the February 2018 Bioinformatics meeting.

eDNA

The purpose of the eDNA Community of Practice is to develop a venue for people interested in the use of eDNA techniques for Earth science in order to improve communication, share knowledge, and catalyze collaboration. eDNA is DNA that is released from an organism into the environment, and that can be used for inventory and monitoring of native and invasive species (Pilliod and others, 2013). In FY 2018, the group's first year as a CDI collaboration area, activities included promotion of eDNA webinars from other venues, gauging interest in different formats and methods for knowledge sharing, and discussion of data releases.

Elevation

The purpose of the Elevation Focus Group is to collate data, information, and tools on both terrestrial and bathymetric elevation topics. Different elevation data types addressed by this group included terrestrial elevation, topo-bathy light detection and ranging (lidar), and bathymetric elevation. The group leaders worked to compile known data sources, provide discussion forums, and make resources more accessible to users through their wiki space.

Subduction Zone

The purpose of the Subduction Zone Focus Group is to exchange information related to advancing Subduction Zone Science (Gomberg and others, 2017) to a broad audience of not only those that are already involved in the research, but others who are interested in learning more. The group lead maintains a wiki page of news, action items, meeting notes, programs, projects, activities, proposal opportunities, and relevant documents such as science plans and reports.

Geographic Information Systems

The purpose of the GIS Community of Practice is to discuss GIS technology, enterprise solutions, and challenges. In FY 2018, the USGS Office of Enterprise Information and the large and active USGS GIS community initiated this group to complement existing GIS resources available to USGS GIS users. The group leaders conducted a survey of community needs and capacity, hosted two widely attended ArcGIS Pro webinars (fig. 8), and promoted the Alaska GIS and Data Science webinars. A topic that the group plans to explore further is open-source GIS-related tools.

Interagency Collaborative for Environmental Modeling and Monitoring

The ICEMM is a U.S. Federal government group chartered through a memorandum of understanding. ICEMM seeks to facilitate collaboration and coordination among Federal agencies in research and development of environmental models and associated monitoring tools, software, and databases. ICEMM supports interagency investigations into environmental risk assessments, uncertainty analyses, and water supply and water quality issues. ICEMM held its annual meeting in April 2018 with the theme “Monitoring and Modeling Data Fusion.” Presentations from the meeting are available on the CDI website (https://my.usgs.gov/confluence/x/0K5tI), and a meeting report was published in EOS (Rashleigh and Nicholson, 2018).

Metadata Reviewers

The purpose of the Metadata Reviewers Community of Practice is to provide a forum for people who review metadata so that consistent standards can be used throughout the USGS. This group enables people new to this role to learn from experienced metadata reviewers. The group met monthly to discuss various topics related to metadata review, listen to presentations, and improve resources for USGS metadata reviewers. Topics in FY 2018 included metadata training needs in the USGS, data quality, a demo of the Alaska Data Integration Working Group (ADIwg) metadata editor (fig. 9), and metadata for specific data types such as genetics data (appendix table 1.6).

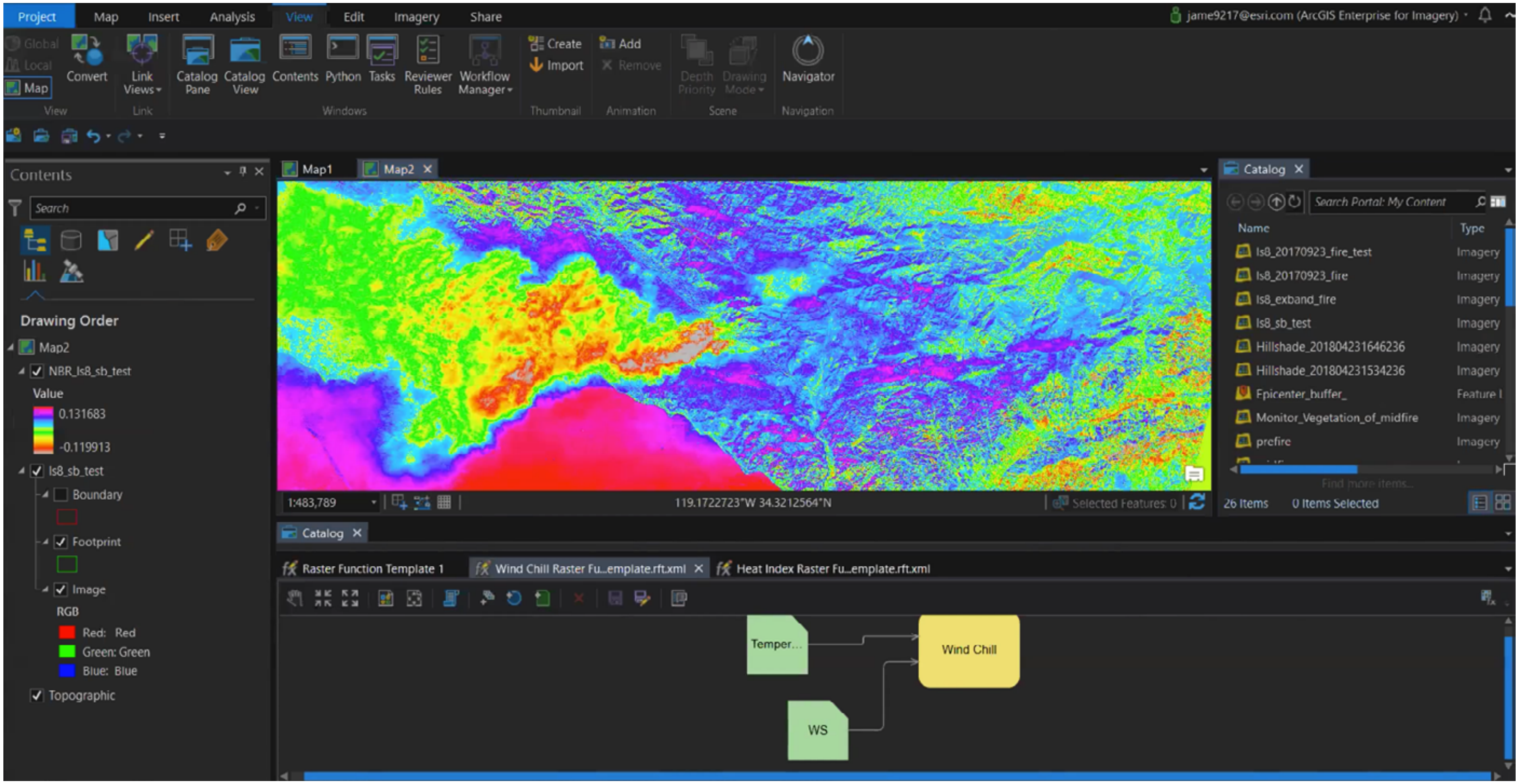
An important result of the Metadata Reviewers Community was recommending revisions to the USGS checklists for review of metadata and data that are used for the review and approval of data releases; the recommended revisions became the basis of the revised checklists approved by the Fundamental Science Practices Advisory Committee and posted on the USGS Data Management website (https://www.usgs.gov/products/data-and-tools/data-management/data-release#checklists). The Metadata Reviewers Community also provided support and volunteers for the 2018 CDI-funded project, “Content Specifications to Enable USGS Transition to ISO Metadata Standard.”

Semantic Web

The Semantic Web Working Group is a group of data practitioners who are working together to explore semantic web technologies to improve the discovery, access, use, and integration of USGS data. In FY 2018, topics included user stories for a permanent USGS triple store (a database for the storage and retrieval of triples through semantic queries, where a triple is a data entity composed of subject-predicate-object), user stories for a USGS database of dictionary elements, a demonstration of the use of permanent identifiers in linked data, and the FAIR data principles (where FAIR stands for Findable, Accessible, Interoperable, and Reusable).

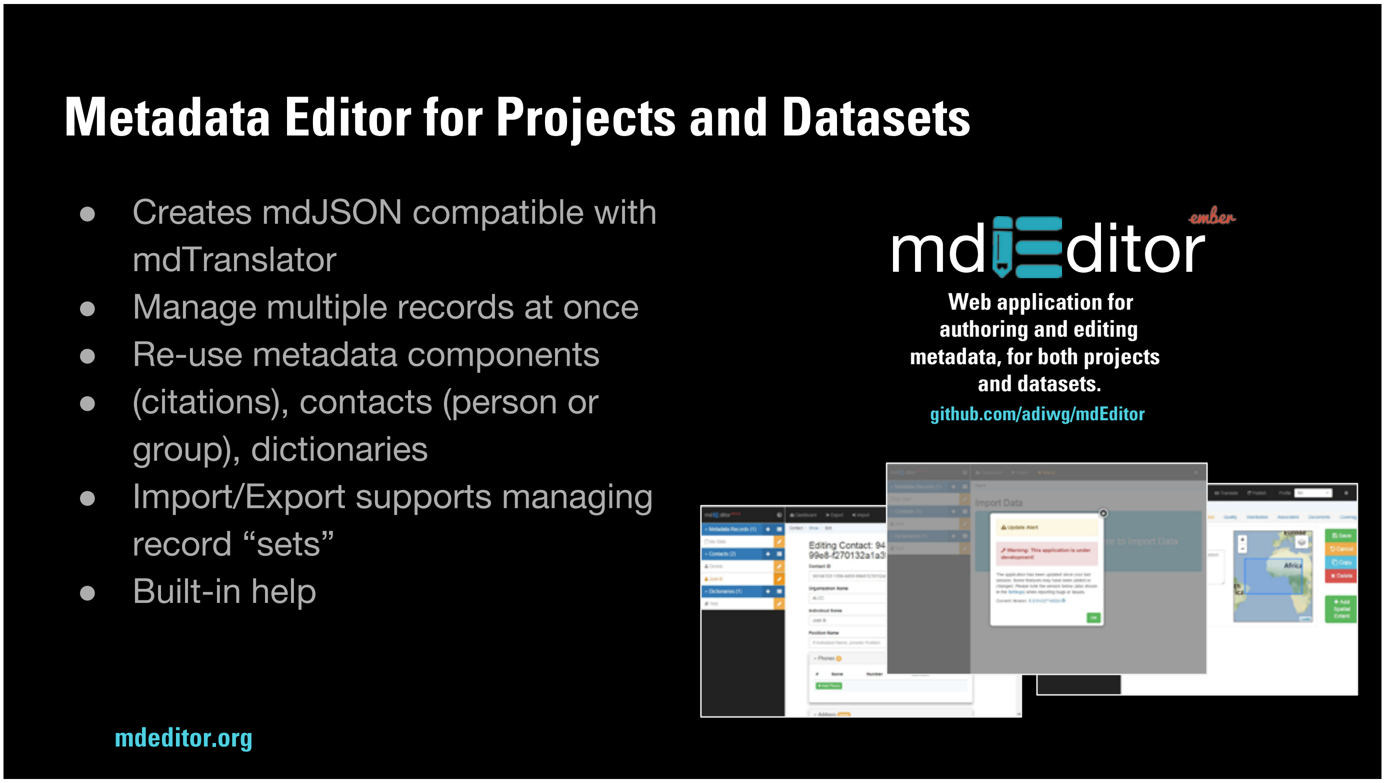
Software Development

The Software Development Cluster is a community for USGS software developers and other interested parties to discuss software release protocols and policies; development of best practices; software metadata; and software libraries, packages, and tools. Topics in FY 2018 included several presentations on different aspects of Git, the USGS EDGE (Equipment and Development Grade Evaluation) program, 508 compliance for websites and web applications, and software licensing aspects (appendix table 1.7).



Name	Type
ls8_20170923_fire_test	Imagery
ls8_20170923_fire	Imagery
ls8_exband_fire	Imagery
ls8_sb_test	Imagery
Hillshade_201804231646236	Imagery
Hillshade_201804231534236	Imagery
Epicenter_buffer_	Feature l
Monitor_Vegetation_of_midfire	Imagery
prefire	Imagery

Figure 8. Image from the April 2018 presentation “Transition from Desktop to ArcGIS Pro” (James Sill, Esri).



Metadata Editor for Projects and Datasets

- Creates mdJSON compatible with mdTranslator
- Manage multiple records at once
- Re-use metadata components
- (citations), contacts (person or group), dictionaries
- Import/Export supports managing record “sets”
- Built-in help

mdEditor ember

Web application for authoring and editing metadata, for both projects and datasets.

github.com/adiwg/mdEditor

mdeditor.org

Figure 9. Slide from the February 2018 presentation “Demo of the New ADIwg Metadata Editor” (Dennis Walworth, U.S. Geological Survey, and Josh Bradley, U.S. Fish and Wildlife Service). (ADIwg, Alaska Data Integration Working Group)

Technology Stack

The goal of the Technology Stack Working Group is to explore and share technologies that aid data discovery, access, and interoperability. The group informs USGS providers and users about tools and techniques to improve efficiency when working with scientific data. The Technology Stack Working Group continued its partnership with Earth Science Information Partners for the Tech Dive webinar series. A focus of FY 2018 was new technologies for scalable computing on massive datasets in the cloud and on high performance computing clusters, using reproducible notebook technologies like JupyterHub. Members were introduced to specific topics such as Jupyter Widgets, the Pangeo project, the Zarr data format, NetCDF-CF advances, and NSF’s Jetstream Cloud (appendix table 1.8).

Special Events and Training

2018 Community for Data Integration Session at the Earth Science Information Partners Summer Meeting

The CDI held a working session for FY 2018 funded projects at the summer Earth Science Information Partners meeting July 17–20, 2018, as its special event in lieu of the biannual CDI workshop. The theme of the session was “Supporting Integrated and Predictive Science: Community for Data Integration Focus on Risk Assessment,” and the purpose of the workshop was to bring together the community to discuss current topics, shared challenges, and steps forward to advance integrated science at the USGS. Six speakers gave progress reports on their FY 2018 funded projects (appendix table 1.9). During the discussion, questions came up about use of existing platforms like GeoPlatform.gov, lessons for working with stakeholders, handling large data volumes and speed of applications, and sustainability of projects.

Python for Data Management

The CDI strives to help its members gain skills in data management and data science that will help people be more efficient in their daily tasks. To this end, the CDI helped to promote a training titled "Python for Data Management." The three-part training series included the broad topics "Working with Local Files," "Batch Metadata Handling," and "Using the USGS ScienceBase Platform with PySB." (PySB is now called sciencebasepy). The training made use of the Jupyter notebook and python environment within the Metadata Wizard 2.0 (Talbert, 2017), which is a successor of the CDI-funded Metadata Wizard (Ignizio and others, 2014). The training was run by Drew Ignizio and Madison Langseth; sponsored by the Science, Analytics, and Synthesis program (at the time, named Core Science Analytics, Synthesis, and Libraries); and attended by 381 unique attendees over the three sessions.

Get Your Science Used

In order to learn insights, tips, and strategies for making products that a targeted audience can understand and use, the CDI hosted a discussion of the USGS circular "Get Your Science Used—Six Guidelines to Improve Your Products" (Perry and others, 2016). Topics that were addressed during this discussion included using plain language at the level of general public literacy, defining goals and audiences, getting honest feedback on the appropriateness of products for the intended audience, communicating difficult topics such as statistical probability, and using the USGS Office of Communications as a resource for evaluation of products.

DataCamp Introduction to Git

The CDI receives frequent inquiries about using Git software for versioning digital files. To attempt to meet this need, we piloted a new training method using free, online self-paced tutorials. Out of several free resources, we chose the DataCamp tutorial Introduction to Git for Data Science (https://www.datacamp.com/courses/introduction-to-git-for-data-science) on recommendation from a CDI member. We facilitated the completion of the online course by sending reminders of weekly assignments to 33 registrants and also facilitated discussion and questions on our wiki space (https://my.usgs.gov/confluence/x/nY0FJ). The course covered topics related to basic workflow, repositories, undo, working with branches, and collaborating. Participants indicated that the accountability of doing the online training in a group with set deadlines was helpful. As a result, we planned additional group learning events for the CDI for the next fiscal year.

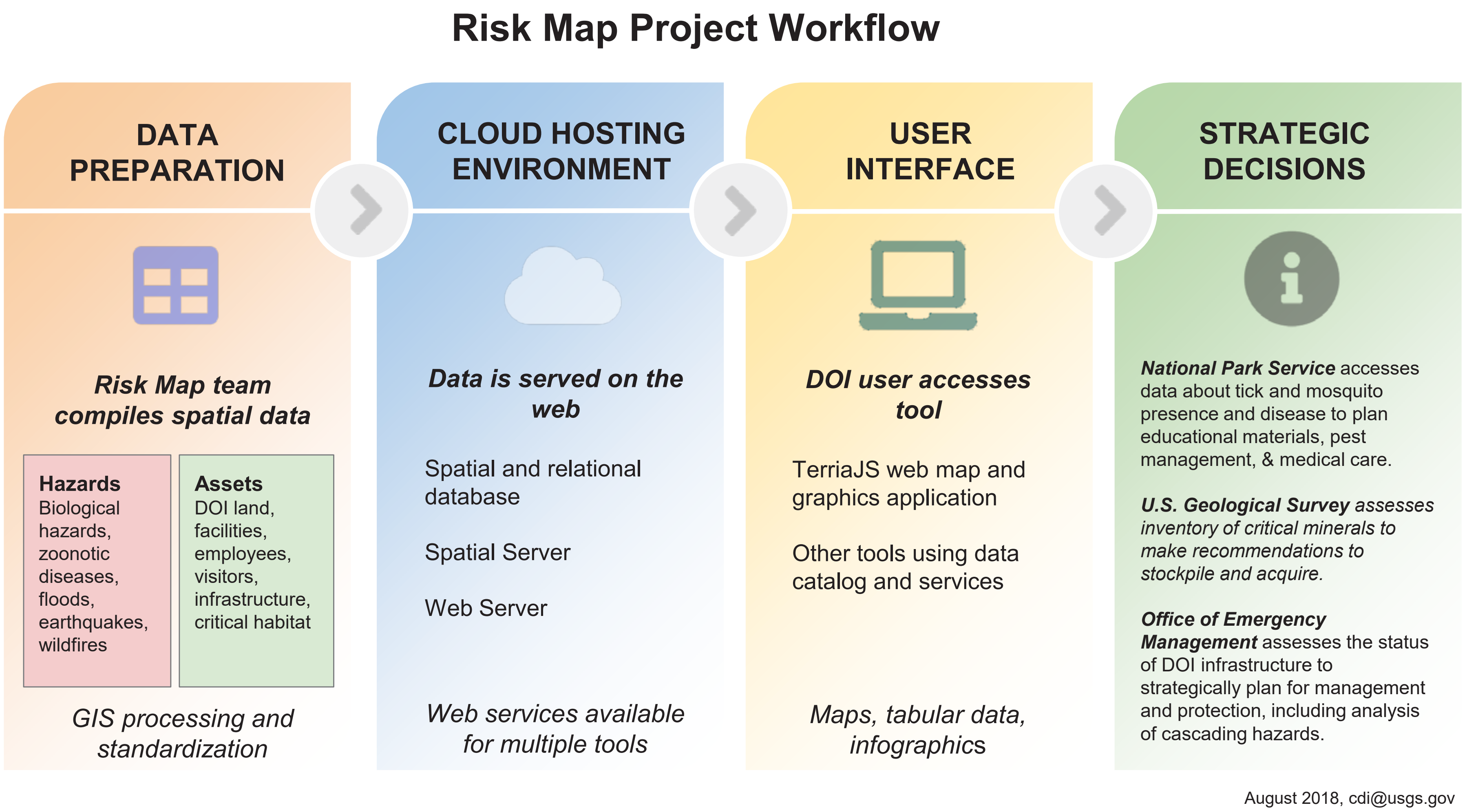
Risk Theme

Community for Data Integration Risk Map Pilot Project

The CDI Risk Map Project (Wood and others, 2018) developed modular tools and services to benefit a wide group of scientists and managers that deal with various aspects of risk research and planning. Risk is the potential that exposure to a hazard will lead to a negative consequence to an asset such as human or natural resources. This project builds upon the Strategic Hazard Identification and Risk Assessment of Department of the Interior Resources project that is developing geospatial layers and other analytical results that visualize multi-hazard exposure to various Department assets (Wood and others, 2019). The CDI Risk Map team developed a spatial database of hazards and assets, an application programming interface (API) to query the data, web services with Geoserver (an open-source geospatial server), and a modular map viewer and related infographics using the open source visualization framework TerriaJS (fig. 10).

Risk Workshop

In January 2018, the CDI organized a workshop that convened 20 participants to discuss select USGS activities related to characterizing, mapping, and communicating societal risk to hazards, with primary focus on the CDI-funded Risk Map Project. During the workshop, participants discussed the CDI Risk Map Project's scope, potential contributions from different entities, and the potential for other CDI-funded projects in FY 2018 that could complement the CDI Risk Map effort. Participants discussed their separate roles in risk research and collaborations that could advance the Integrated Predictive Science Capacity described in the FY 2018 USGS Bureau Priorities (USGS, 2017b).



Risk Map Project Workflow

DATA PREPARATION

Risk Map team compiles spatial data

Hazards
Biological hazards, zoonotic diseases, floods, earthquakes, wildfires

Assets
DOI land, facilities, employees, visitors, infrastructure, critical habitat

GIS processing and standardization

CLOUD HOSTING ENVIRONMENT

Data is served on the web

Spatial and relational database

Spatial Server

Web Server

Web services available for multiple tools

USER INTERFACE

DOI user accesses tool

TerriaJS web map and graphics application

Other tools using data catalog and services

Maps, tabular data, infographics

STRATEGIC DECISIONS

National Park Service accesses data about tick and mosquito presence and disease to plan educational materials, pest management, & medical care.

***U.S. Geological Survey** assesses inventory of critical minerals to make recommendations to stockpile and acquire.*

Office of Emergency Management assesses the status of DOI infrastructure to strategically plan for management and protection, including analysis of cascading hazards.

August 2018, cdi@usgs.gov

Figure 10. Schematic of the Community for Data Integration risk map project workflow from data preparation to strategic decisions. (DOI, U.S. Department of the Interior; GIS, geographic information system)

Outputs from the workshop illustrate the CDI's goal of increasing communication and collaboration within the USGS. Outputs include (1) establishment of a community interested in the CDI Risk Map project that goes beyond the CDI; (2) identification of additional potential cross-effort relationships and shared contributions to the CDI risk Map project; (3) identification of ways that the groups represented by the workshop participants could work together to augment the CDI Risk Map Project; (4) links between existing program data assets, methods, and capabilities; (5) links between the USGS Plan for Risk Research and Applications, "Science for a Risky World: A USGS Plan for Risk Research & Applications" (Ludwig and others, 2018), and the CDI to build a risk community of practice; and (6) an inventory of several USGS risk mapping efforts, which are summarized in table 2.

Table 2. Selected U.S. Geological Survey efforts in risk research.

[CDI, Community for Data Integration; USGS, U.S. Geological Survey; DOI, U.S. Department of the Interior; GIS, geographic information system; CONUS, continental United States]

USGS effort	Strategic direction provided by	Description
CDI Risk Map	Community for Data Integration	Reusable and modular tools and data, examples of building blocks for Integrated Predictive Science Capacity
USGS Risk Plan	Natural Hazards Mission Area	Strategic plan for risk research and applications (Ludwig and others, 2018)
Strategic Hazard Identification and Risk Assessment on DOI Resources Project (SHIRA)	DOI Office of Emergency Management	GIS layers and related analytical products for the Strategic Hazard Identification and Risk Assessment on DOI Resources that can be incorporated into web applications based on stakeholder input at a February 2018 workshop (Wood and others, 2019)
Core Science Systems Mission Area Risk Map Support	Core Science Systems Mission Area	Opportunities to use Core Science Analytics, Synthesis, and Library[1] expertise for supporting DOI Risk Map and (or) CDI Risk Map effort

[1] The name of the Core Science Analytics, Synthesis, and Library has been changed to Science Analytics and Synthesis.

Community for Data Integration Request for Proposals Theme

In FY 2018, the CDI Executive Sponsors encouraged proposals that would produce building blocks for an Integrated Predictive Science Capacity in the specific focus area of risk assessment and hazard exposure. Integrated Predictive Science Capacity was described in the USGS FY 2018 Bureau Priorities as "powerful new products and services that provide: (1) vulnerability detection and assessment, (2) prediction and forecasting, (3) early warning, and (4) decision support at the scale of decisions." Examples of risk-related projects include integration of models and data to forecast invasive species, developing methods for aligning and integrating risk and hazard data from different sources, and technology to improve information delivery and stakeholder feedback in risk and hazard activities (USGS, 2017b). More detail is included in the following section.

Annual Community for Data Integration Request for Proposals

The CDI annually supports innovative projects that produce new and reusable ideas, methods, or tools that have an impact beyond a single USGS program, center, region, or mission area. The CDI provides up to $50,000 per project. Project proposals are evaluated based on (1) their alignment with the CDI Science Support Framework (USGS, 2014); (2) the evaluation criteria laid out in the request for proposals guidance document (USGS, 2017) including scope, technical approach, project experience and collaboration, sustainability, budget justification, and timeline; and (3) how the proposal supports the following CDI guiding principles (USGS, 2017):

- focus on targeted efforts that yield near-term benefits to Earth and biological science;
- leverage existing capabilities and data;
- implement and demonstrate innovative solutions, such as methodologies, tools, or integration concepts, that could be used or replicated by others at scales from project to enterprise;

- preserve, expose, and improve access to Earth and biological science data, models, and other outputs; and
- develop, organize, and share knowledge and best practices in data integration.

Following the request for proposals, 32 statements of interest were submitted. The principal investigators and collaborators on the statements of interest represented six USGS mission areas that, at the time, were named Climate and Land-Use Change, Ecosystems, Natural Hazards, Water, Core Science Systems, and Energy and Minerals. Of the submitted statements, 16 addressed the risk theme, with topics including forecasting invasive species, floodplain modeling, wildfire risk prediction, landslides, and drought risk (https://my.usgs.gov/confluence/x/SwL8Ig).

The "Community for Data Integration Projects" section provides a summary of each of the nine projects funded in FY 2018 under the request for proposals process (table 3). A description of the accomplishments for each of the projects will be provided in a separate report.

Table 3. Overview of the Community for Data Integration projects funded in fiscal year 2018 (in alphabetical order by principal investigator last name). Project title hyperlinks resolve to a ScienceBase web page describing the project and linking to external resources such as publications, code repositories, and related websites.

[lidar, light detection and ranging; USGS, U.S. Geological Survey]

Title	URL	Lead principal investigator(s)	Lead program
"An Interactive Web-Based Tool for Anticipating Long-Term Drought Risk"	https://www.sciencebase.gov/catalog/item/5acd21aae4b0e2c2dd155dea	John Bradford	Southwest Biological Science Center
"ICE! Ice Jam Hazard Mobile-Friendly Website"	https://www.sciencebase.gov/catalog/item/5b9198e5e4b0702d0e808b76	Katherine Chase	Montana-Wyoming Water Science Center
"National Alert Risk Mapper for Non-indigenous Aquatic Species"	https://www.sciencebase.gov/catalog/item/5acd257ae4b0e2c2dd155df5	Pam Fuller	Southeast Ecological Science Center
"Integrating Disparate Spatial Datasets from Local to National Scale for Open-Access Web-Based Visualization and Analysis: A Case Study Compiling U.S. Landslide Inventories"	https://www.sciencebase.gov/catalog/item/5acd2600e4b0e2c2dd155dfa	Benjamin Mirus	Geologic Hazards Science Center
"Knowledge Extraction Algorithms (KEA): Turning Literature Into Data"	https://www.sciencebase.gov/catalog/item/5acd2680e4b0e2c2dd155dfd	Matthew Neilson	Southeast Ecological Science Center
"Investigation of Lidar Data Processing and Analysis in the Cloud"	https://www.sciencebase.gov/catalog/item/5b919e87e4b0702d0e808b9d	Jessica Walker	Western Geographic Science Center
"Content Specifications to Enable USGS Transition to ISO Metadata Standard"	https://www.sciencebase.gov/catalog/item/5acd27a0e4b0e2c2dd155e01	Dennis Walworth	Alaska Science Center
"Mapping Land-Use, Hazard Vulnerability and Habitat Suitability Using Deep Neural Networks"	https://www.sciencebase.gov/catalog/item/5acd2923e4b0e2c2dd155e09	Jonathan Warrick	Pacific Coastal and Marine Science Center
"Workflows to Support Integrated Predictive Science Capacity: Forecasting Invasive Species for Natural Resource Planning and Risk Assessment"	https://www.sciencebase.gov/catalog/item/5acd27b3e4b0e2c2dd155e03	Jake Weltzin	National Phenology Network

Community for Data Integration Projects

The following sections provide the plain language summaries of the FY 2018 CDI request for proposals proposal for each project.

An Interactive Web-Based Tool for Anticipating Long-Term Drought Risk

Droughts are becoming more frequent and severe and this trend is expected to continue in coming decades. Drought effects on natural resources include reduced water availability for plants and humans; increased insect, disease, and fire outbreaks; and increased vegetation mortality. Although recognition of drought risk is growing, no tools are available to help natural resource managers understand site-specific potential exposure to future droughts. Managers need long-term historical perspectives so they can recognize their site's natural range of variability. Managers also need future projections of 21st century drought in order to quantify their site's risk of drought exposure. We propose to develop an online, interactive platform that helps users access and visualize site-specific, historical, and future water availability. Users can identify a location, and (optionally) specify soil and vegetation conditions. Using these inputs, our platform will estimate past and future drought conditions, and deliver synthesized and raw data for analysis.
Contact: John Bradford, Southwest Biological Science Center, jbradford@usgs.gov

ICE! Ice Jam Hazard Mobile-Friendly Website

Ice jams along rivers cause flooding, scouring, human injuries and loss of life, and structural and environmental damage; they are a major hazard across the northern United States. Communities need data about ice jam locations and frequencies, as well as information about developing ice jams that might threaten lives and property. This project will enable individuals to collect real-time information about ice jams. We plan to test the mobile-friendly website in Montana (and potentially other states), and then make it available throughout the United States. Ice jam data are planned to be available to the public on online maps and formatted for transfer to the U.S. Army Corps of Engineers Ice Jam Archive. The maps and archive will be useful for understanding ice jam processes and identifying sites that are vulnerable to ice jam flooding and damages. The data can be used in tools to help predict the probability of ice jams across the Nation.
Contact: Katherine Chase, Montana-Wyoming Water Science Center, kchase@usgs.gov

National Alert Risk Mapper for Nonindigenous Aquatic Species

The Nonindigenous Aquatic Species (NAS) Database and Alert System provides a framework for the rapid dissemination of information about new invasions. The system notifies registered users of new sightings of more than 1,270 nonnative species as part of national-scale early detection and rapid response systems. The NAS group is developing a new tool, the NAS Alert Risk Mapper to characterize water bodies potentially at risk from a new nonnative species sighting. We propose to improve the risk mapper's effectiveness as a tool for the early detection and rapid response systems by expanding its geographical extent to include the entire contiguous United States and Hawaii and automating the map making process. Maps from the NAS Alert Risk Mapper will indicate lakes and river reaches that are at risk of invasion from a new nonindigenous species sighting within drainages and will accompany the NAS Alert emails sent out to subscribers and managers.
Contact: Pam Fuller, Southeast Ecological Science Center, pfuller@usgs.gov

Integrating Disparate Spatial Datasets from Local to National Scale for Open-Access Web-Based Visualization and Analysis—A Case Study Compiling U.S. Landslide Inventories

Understanding spatial patterns is fundamental to Earth science and risk assessments, but spatial data are often collected at local scales, in disparate formats, and within specific jurisdictional boundaries. For example, weather-triggered landslides pose a major threat to public safety and economic well-being across the United States, but detailed information on their occurrence and susceptibility varies greatly in data quality, accessibility, and spatial extent. A goal of this project is to compile existing data on landslide occurrence within the United States into a searchable, web-based map for use by the public, researchers, and emergency planners. The project seeks broader input from the CDI community to ultimately provide general guidance, workflows, and sustainable data management practices for integrating local-scale spatial data into national-scale visualization and analysis products. The dual benefits would be to enhance access to landslide information and facilitate future integration of other interdisciplinary spatial data into national-scale risk assessment products.
Contact: Benjamin Mirus, Geologic Hazards Science Center, bbmirus@usgs.gov

Knowledge Extraction Algorithms—Turning Literature Into Data

Research efforts often rely on identification and extraction of information from published literature. The USGS NAS database uses literature to help populate the database with records of species occurring outside their native range. Currently, this information is extracted from literature manually, which is time consuming (for example, reading articles, typing data from literature into database and more), introduces potential data entry mistakes, and results in records that are difficult to reproduce. New technologies are being explored in the USGS to programmatically identify and extract relevant information from digital documents. This project will develop computer code to identify and extract relevant information from the literature, helping make the update of NAS more efficient and sustainable. An overarching goal of the effort is to test the viability of building a systematic method for knowledge extraction from literature.
Contact: Matthew Neilson, Southeast Ecological Science Center, mneilson@usgs.gov

Investigation of Lidar Data Processing and Analysis in the Cloud

The multiagency 3DEP project addresses the growing demand for detailed topography and surface models through the systematic acquisition of high-quality lidar data. Currently, users download 3DEP products to local computers for processing and analysis. Massive data volumes, limited transfer rates, and the restricted power of desktop systems hamper both accessibility and efficiency. This project investigates the potential for facilitating data derivation and analysis by leveraging cloud-computing functionality and lidar-specific indexing software. We plan to demonstrate how common extraction and analysis commands can be performed on lidar datasets hosted in the Amazon Web Services environment through a local web browser. The 3DEP project identifies a representative use-case dataset, determines appropriate open-source software, and develops the analysis methodology, which will be implemented and tested by USGS Cloud Hosting Solutions personnel. The result will be a generalizable workflow that can be scaled to meet diverse lidar application needs across the USGS.
Contact: Jessica Walker, Western Geographic Science Center, jjwalker@usgs.gov

Content Specifications to Enable USGS Transition to ISO Metadata Standard

USGS has been a leader in producing high quality metadata records to enable discovery and reuse of research data. (Numerous examples may be found in the USGS Science Data Catalog at https://data.usgs.gov/datacatalog/.) In the 1990s, USGS embraced the Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata and developed tools and quality-control systems for use across the bureau. Soon it will be time for USGS to transition to the international metadata standards known collectively as ISO 19115. As a step toward the transition, this project is convening a workshop of data specialists from across the bureau to propose content specifications for USGS to aid in authoring metadata records in the new standard. These specifications will enable USGS to maintain metadata quality while also enjoying the flexibility and modern features of the ISO standard. The project aims to present a proof of concept, by using the proposed specifications in the metadata toolkit developed by the interagency Alaska Data Integration workgroup (https://mdtools.adiwg.org/). On completion of the project, the new specifications will be offered to the USGS CDI for review and improvement. The procedures from the workshop will also be published for future use to develop additional specifications that will be needed to fit the diverse types of research data that USGS produces.
Contact: Dennis Walworth, Alaska Science Center, dwalworth@usgs.gov

Mapping Land-Use, Hazard Vulnerability, and Habitat Suitability Using Deep Neural Networks

Deep learning is a computer analysis technique inspired by the human brain's ability to learn. It involves several layers of artificial neural networks to learn and subsequently recognize patterns in data, forming the basis of many state-of-the-art applications from self-driving cars to drug discovery and cancer detection. Deep learning represents a paradigm shift in the analysis of large, complicated data, and is the fastest-growing trend in the analysis of physical, geological, and biological data. Deep neural networks are capable of learning many levels of abstraction, and thus outperform many other types of automated classification algorithms. This project aims to develop software tools, resources, and training workshops that will allow USGS scientists to apply deep learning to remotely sensed imagery and to better understand natural hazards and habitats across the Nation.
Contact: Jonathan Warrick, Pacific Coastal and Marine Science Center, jwarrick@usgs.gov

Workflows to Support Integrated Predictive Science Capacity—Forecasting Invasive Species for Natural Resource Planning and Risk Assessment

Timely information regarding the risk of activity of invasive plant and animal species, which are biological hazards, is critical for efficient detection and control efforts by resource managers. The USA National Phenology Network proposed to produce and deliver national-scale, real-time, and short-term forecast maps indicating the optimal time to detect and treat problematic, invasive insect pests including emerald ash borer and hemlock wooly adelgid. These risk maps aim to enhance decision making and short-term planning by both natural resource managers and members of the public. The project plans to create workflows and modular, transferable tools and services to benefit USGS and associated stakeholder communities. A key goal is improved capacity within USGS for ecological forecasting, risk assessment, and translational (coproduced) science related to biological hazards. Key stakeholders include natural resource managers within protected areas, such as national parks and wildlife refuges, and within lands managed by municipalities, states, tribes and private landholders.
Contact: Jake Weltzin, National Phenology Network, jweltzin@usgs.gov

Summary—Building an Integrated Predictive Science Capacity

Through the monthly forums, workshops, working groups, projects, and constant surveying of the community's needs, the Community for Data Integration (CDI) provides content that keeps members informed of information and tools to work with their data. The community continues to grow in size and scope, with new collaboration areas forming in Citizen-Centered Innovation, eDNA (Environmental DNA), Elevation, Geographic Information Systems, and Subduction Zone Science. In FY 2018, the CDI tested new learning methods such as online group learning and reading discussion groups. The CDI has made some changes to make its outputs more discoverable and usable. These include separating out the funded project final reports from the annual report (Hsu and others, 2018a) and reviewing and standardizing our output categories and making them programmatically searchable on ScienceBase (https://www.sciencebase.gov/catalog/item/520e8340e4b08494c3cb34ec).

CDI's sponsors have challenged the community to use its expertise to help build a U.S. Geological Survey integrated predictive science capacity. By focusing the funded part of our activities on the risk theme, pilot project "CDI Risk Map" in FY 2018, and associated coordination activities around the risk theme, the CDI facilitation team focused its efforts on a demonstrable capability of Integrated Predictive Science Capacity. The risk theme focus demonstrated the utility of standardizing data and documenting workflows for use among multiple project teams. As we increase in visibility at the U.S. Geological Survey and beyond, we will continue to facilitate activities to support data and science integration activities for the Earth and biological sciences.

References Cited

Gomberg, J.S., Ludwig, K.A., Bekins, B.A., Brocher, T.M., Brock, J.C., Brothers, D., Chaytor, J.D., Frankel, A.D., Geist, E.L., Haney, M., Hickman, S.H., Leith, W.S., Roeloffs, E.A., Schulz, W.H., Sisson, T.W., Wallace, K., Watt, J.T., and Wein, A., 2017, Reducing risk where tectonic plates collide—U.S. Geological Survey subduction zone science plan: U.S. Geological Survey Circular 1428, 45 p., accessed October 21, 2019, at https://doi.org/10.3133/cir1428.

Hsu, L., Allstadt, K.E., Bell, T.M., Boydston, E.E., Erickson, R.A., Everette, A.L., Lentz, E., Peters, J., Reichert, B.E., Nagorsen, S., Sherba, J.T., Signell, R.P., Wiltermuth, M.T., and Young, J.A., 2018a, Community for Data Integration fiscal year 2017 funded project report: U.S. Geological Survey Open-File Report 2018–1154, 15 p., accessed October 21, 2019, at https://doi.org/10.3133/ofr20181154.

Hsu, L., Hutchison, V.B., Langseth, M.L., and Wheeler, B., 2018b, U.S. Geological Survey Community for Data Integration 2017 Workshop Proceedings: U.S. Geological Survey Open-File Report 2018–1081, 56 p., accessed October 21, 2019, at https://doi.org/10.3133/ofr20181081.

Ignizio, D.A., O'Donnell, M.S., and Talbert, C.B., 2014, Metadata wizard—An easy-to-use tool for creating FGDC–CSDGM metadata for geospatial datasets in Esri ArcDesktop: U.S. Geological Survey Open-File Report, 2014–1132, 14 p., accessed October 21, 2019, at https://doi.org/10.3133/ofr20141132.

Kellogg, C.A., and Hopkins, M.C., 2017, USGS microbiome research: U.S. Geological Survey Fact Sheet 2017–3074, 4 p., accessed October 21, 2019, at https://doi.org/10.3133/fs20173074.

Ludwig, K.A., Ramsey, D.W., Wood, N.J., Pennaz, A.B., Godt, J.W., Plant, N.G., Luco, N., Koenig, T.A., Hudnut, K.W., Davis, D.K., and Bright, P.R., 2018, Science for a risky world—A U.S. Geological Survey plan for risk research and applications: U.S. Geological Survey Circular 1444, 57 p., accessed October 21, 2019, at https://doi.org/10.3133/cir1444.

Office of Science & Technology Policy, 2019, Implementation of Federal prize and citizen science authority: fiscal years 2017–18, accessed October 21, 2019, at https://www.whitehouse.gov/wp-content/uploads/2019/06/Federal-Prize-and-Citizen-Science-Implementation-FY17-18-Report-June-2019.pdf.

Perry, S.C., Blanpied, M.L., Burkett, E.R., Campbell, N.M., Carlson, A., Cox, D.A., Driedger, C.L., Eisenman, D.P., Fox-Glassman, K.T., Hoffman, S., Hoffman, S.M., Jaiswal, K.S., Jones, L.M., Luco, N., Marx, S.M., McGowan, S.M., Mileti, D.S., Moschetti, M.P., Ozman, D., Pastor, E., Petersen, M.D., Porter, K.A., Ramsey, D.W., Ritchie, L.A., Fitzpatrick, J.K., Rukstales, K.S., Sellnow, T.S., Vaughon, W.L., Wald, D.J., Wald, L.A., Wein, A., and Zarcadoolas, C., 2016, Get your science used—Six guidelines to improve your products: U.S. Geological Survey Circular 1419, 37 p., accessed October 21, 2019, at https://doi.org/10.3133/cir1419.

Pilliod, D.S., Goldberg, C.S., Laramie, M.B., and Waits, L.P., 2013, Application of environmental DNA for inventory and monitoring of aquatic species: U.S. Geological Survey Fact Sheet 2012–3146, 4 p.

Rashleigh, B., and Nicholson, T., 2018, Agencies collaborate to better monitor and model the environment: Eos—Earth & Space Science News web page, accessed October 21, 2019, at https://doi.org/10.1029/2018EO106651.

Talbert, C., 2017, MetadataWizard: U.S. Geological Survey, accessed October 21, 2019, at https://doi.org/10.5066/f7v9870d.

U.S. Geological Survey [USGS], 2014, CDI Science Support Framework: U.S. Geological Survey web page, accessed February 10, 2019, at https://my.usgs.gov/confluence/display/cdi/CDI+Science+Support+Framework.

U.S. Geological Survey [USGS], 2017a, U.S. Geological Survey (USGS) Community for Data Integration (CDI) request for proposals (RFP) for fiscal year 2018: U.S. Geological Survey, 18 p., accessed February 10, 2019, at https://my.usgs.gov/confluence/display/cdi/2018+Proposals?preview=/559842450/583664750/CDI_FY18_Request_for_Proposals_final.pdf#id-2018Proposals-GuidanceDocument.

U.S. Geological Survey [USGS], 2017b, USGS Director's annual bureau guidance for fiscal year 2018: U.S. Geological Survey Memorandum to all U.S. Geological Survey employees from William Werkheiser, Acting Director, June 9, 2017, accessed October 21, 2019, at https://my.usgs.gov/confluence/x/fb4RIg.

Wood, N.J., Jones, J.M., Henry, K.D., Sherba, J.T., Ng, P., CDI Risk Map, 2018: U.S. Geological Survey ScienceBase data release, accessed August 2019 at https://www.sciencebase.gov/catalog/item/5b91a0c2e4b0702d0e808bb2.

Wood, N., Pennaz, A., Ludwig, K., Jones, J., Henry, K, Sherba, J., Ng, P., Marineau, J., and Juskie, J., 2019, Assessing hazards and risks at the Department of the Interior—A workshop report: U.S. Geological Survey Circular 1453, 42 p., accessed October 21, 2019, at https://doi.org/10.3133/cir1453.

Glossary

Cluster A less formal group that can range from a mailing list to a group that maintains online collections of resources on their area of interest.

Collaboration area A group formed around a specific topic that provides a platform for sharing resources and knowledge, discussing challenges, and identifying solutions that will help to advance data integration in the Earth and biological sciences.

Community of practice A group of people who share a concern or a passion for something they do and learn how to do it better as they interact regularly (https://wenger-trayner.com/introduction-to-communities-of-practice/). The Community for Data Integration itself is a community of practice, and some of its collaboration areas identify themselves as communities of practice.

Focus group Smaller and usually time-limited groups that form to address more specific issues and report up to their respective collaboration area.

Working group A subgroup of the Community for Data Integration that forms around common interests, helps address challenges, and identifies solutions that enable data integration efforts.

Appendix 1. Presentations and Speakers

This appendix lists presentation topics and speakers in the different forums and collaboration area meetings for fiscal year 2018.

Table 1.1. Monthly Community for Data Integration (CDI) forum presentations for fiscal year (FY) 2018.

[USGS, U.S. Geological Survey; SECORRA/IOOS, Southeast Coastal Ocean Observing Regional Association/Integrated Ocean Observing System; API, application programming interface; lidar, light detection and ranging]

Date	Presentation title	Speaker(s)	Number of Attendees
October 12, 2017	"Reproducible Notebook Series"	Rich Signell, USGS	101
	"From API to Apps: USGS Texas Water Science Center Web Development Approaches and Analytics"	Daniel Pearson, USGS	
November 8, 2017	"Reproducible Notebook Series: Notebooks as a Data Management Superpower"	Colin Talbert, USGS	83
	"National Earthquake Information Center"	Michell Guy and Lynda Lastowka, USGS	
January 10, 2018	"How the Freedom of Information Act Impacts Data"	Brian May, USGS	103
	"The CDI Request for Proposals Process: Past, Present, and Future"	Leslie Hsu, USGS	
February 14, 2018	"Reproducible Notebook Series"	Kyle Enns and Cristina Falvo, USGS	89
	"Semantic Web for Sustainability: Revolutionizing How We Write, Find, Link and Reuse Data and Models"	Ferdinando Villa, Basque Centre for Climate Change (BC3)	
March 14, 2018	"USGS Risk Plan"	Kristin Ludwig, USGS	87
	"An Interactive Web-based Application for Earthquake-Triggered Ground Failure Inventories"	Kate Allstadt, USGS	
	"Flocks of a Feather Dock Together: Using Docker and HTCondor to Link High-Throughput Computing Across the USGS"	Richard Erickson, USGS	
	"Visualizing Community Exposure and Evacuation Potential to Tsunami Hazards Using anIinteractive Tableau Dashboard"	Jeff Peters, USGS	
	"Exploring the USGS Science Data Life Cycle in the Cloud"	Rich Signell, USGS	
April 11, 2018	"Reproducible Notebook Series—OBIS (Ocean Biogeographic Information System) and R"	Filipe Fernandes, SECORRA/IOOS	91
	"Taxa Taxi: An Automated Process for Using Citizen Science Data to Facilitate Biodiversity Monitoring"	Erin Boydston and Toni Lyn Morelli, USGS	
	"USGS Data at Risk: Expanding Legacy Data Inventory and Preservation Strategies"	Lance Everette and Tara Bell, USGS	
	"Web Mapping Application for a Historical Geologic Field Photo Collection"	Sarah Nagorsen, USGS	

Table 1.1. Monthly Community for Data Integration (CDI) forum presentations for fiscal year (FY) 2018.—Continued

[USGS, U.S. Geological Survey; SECORRA/IOOS, Southeast Coastal Ocean Observing Regional Association/Integrated Ocean Observing System; API, application programming interface; lidar, light detection and ranging]

Date	Presentation title	Speaker(s)	Number of Attendees
May 9, 2018	"Developing APIs to Support Enterprise Level Monitoring Using Existing Tools"	Brian Reichert and Becca Scully, USGS	90
	"Extending ScienceCache—a Mobile Application for Data Collection—to Accommodate Broader Use within USGS"	Mark Wiltermuth, USGS	
	"Evaluation and Testing of Standardized Forest Vegetation Metrics Derived from Lidar Data"	John Young, USGS	
June 13, 2018	"Amplifying USGS Science with Timely and Digestible Data Visualizations"	Jordan Read, USGS	129
	"Data Visualization for Science: Comparing 3 Dashboard Building Software Packages"	Kevin Henry, Jason Sherba, and Jeff Peters, USGS	
July 11, 2018	"STEP-UP to Support Students & Science"	Chris Hammond, USGS	89
	"USGS Cloud Hosting Solutions Update"	Jennifer Erxleben and Harry House, USGS	
August 8, 2018	"Briefing on the Component Architecture for Integrated Science"	Tim Quinn, USGS	114
	"Large Data Storage and Access at the USGS"	Nancy Sternberg, USGS	
September 12, 2018	"STEP-UP to Data Management"	Sue Kemp, USGS	86
	"CDI Funded Project Report: Empowering Decision-Makers: A Dynamic Web Interface for Running Bayesian Networks"	Erika Lentz, USGS	
	"USGS Thesaurus: What It Is, How You Can Use It, and How You Can Improve It"	Peter Schweitzer, USGS	

Table 1.2. Citizen-Centered Innovation Community meetings and presentations for fiscal year 2018.

[DOI, Department of the Interior]

Date	Subject	Speaker
February 21, 2018	Overview of "Citizen-Centered Innovation Community," introduction of participants, discussion on "DOI Generic Information Collection Request"	Sophia Liu, USGS
March 21, 2018	Overview of "Citizen-Centered Innovation Community," introduction of participants, continued discussion on "DOI Generic Information Collection Request"	Sophia Liu, USGS
August 15, 2018	Overview of the Report to Congress for the "Crowdsourcing and Citizen Science Act"	Sophia Liu, USGS

Table 1.3. Data Management Working Group monthly meeting presentations for fiscal year 2018.

[NCCWSC, National Climate Change and Wildlife Science Center; USGS, U.S. Geological Survey; FAIR, Findable, Accessible, Interoperable, and Reusable; API, application programming interface; SDC-CMS, Science Data Catalog-content management system]

Date	Title	Speaker
October 16, 2017	"Managing Data with Partners" "Data Management Planning in NCCWSC"	Donn Holmes, USGS Emily Fort, USGS
November 13, 2017	"Guidance on How to Release USGS Model Output Files"	Fran Lightsom, USGS
	"Examples of Building Data Management Plans as Code"	Sky Bristol, USGS
	"Data Management Activities in the Water Mission Area"	Linda Debrewer, USGS
December 11, 2017	"Data Integration, Fiscal Accountability, and the 'Business of Science'"	Brian Reece, USGS
January 8, 2018	"Publishing Metadata to the Science Data Catalog"	Lisa Zolly, USGS
	"Data Management Challenges"	Cassandra Ladino, USGS
February 12, 2018	"Tidy Data—Using Python and Pandas to Clean Legacy Datasets"	Emily Baker, USGS
	"Documenting Scientific Workflows and Findings in Biological Analysis Packages"	Daniel Wieferich, USGS
	The Intersection between Volunteered Geographic Information (VGI) and Federal Geospatial Data	Elizabeth McCartney and Erin Korris, USGS
March 12, 2018	"Capturing Your Processing and Analysis Workflow in R"	Alison Appling, USGS
April 9, 2018	"Taking Action Against USGS Legacy Data Challenges"	Tara Bell and Lance Everette, USGS
May 14, 2018	Quick update on "Data Sharing Agreements"	JC Nelson, USGS
	"Data Management to Support Integrative, FAIR, Multidisciplinary Modeling: Lessons from the Last Decade and Paths Forward"	Ken Bagstad, USGS
June 11, 2018	"Data Source Authority Files"	Drew Ignizio, USGS
	"The Department of the Interior Metadata Implementation Guide"	Ray Obuch, USGS
July 9, 2018	Brief update on "Science Data Catalog and data.gov"	Ben Wheeler, USGS
	Overview and demo of the "Metadata Wizard, a Tool for Creating Robust Metadata"	Colin Talbert, USGS
August 13, 2018	"Relationship Between Records Management and Science Data"	Chris Bartlett, USGS
	"Retiring Scientific Records: Inactive Records, ARCIS, and the Federal Records Center"	Larry Reedy, USGS
September 10, 2018	"Science Data Catalog API Ingest Service (SDC-CMS): Overview for USGS Data Stewards"	Lance Everette, USGS
	"R2O, Research to Operation: Towards a USGS Model"	Ra'ad Saleh, USGS

Table 1.4a. Software Development and Information Technology Operations (DevOps) Project Management Sync topics for fiscal year 2018.

[USGS, U.S. Geological Survey; CHS, Cloud Hosting Solutions; CDN/WAF, content delivery network/web application firewall; GHSC, Geologic Hazards Science Center]

Date	Title	Speaker
October 3, 2017	"Open Shift Demo"	Chuck Svoboda, OpenShift
November 7, 2017	"Intro USGS Git Hosting Platform"	Eric Martinez, USGS
	"CHS CDN/WAF Service"	Jonathan Russo, USGS
December 5, 2017	"SCAPE: A Framework for Adaptable and Secure Analysis of Streaming Data"	Ginny Cevasco, Booz Allen Hamilton
	"GHSC Experience with an Agile Contract"	Lynda Lastowka, USGS
February 6, 2018	"CHS Update"	George Rolston, USGS
	"Agile and DevOps at Booz Allen Hamilton"	Ginny Cevasco, Booz Allen Hamilton
	"Apps Dynamics Product Demo"	Justin Boyle, Apps Dynamics
March 6, 2018	"Cloud Training"	Brian Fox, USGS
	"CHS Update"	Ross Wickman, USGS
	"Software Inventory, What It Is, How It's Made, and How You Can Make It Better"	Eric Martinez, USGS
April 3, 2018	"Software Management Website Update"	Cassandra Ladino, USGS
	"Zero Trust Networks"	Tom Van Dreser, USGS
	"Overview of Cloud Activities at Cal Poly"	Paul Jurasin, California Polytechnic State University
June 5, 2018	"Software Management Website Update"	Cassandra Ladino, USGS
	"A World Wind Tour of cloud.gov and the Default DevOps Pipeline to Deploy Applications to It"	Andrew Burnes, USGS
August 7, 2018	"Community for Data Integration (CDI) Software Development Cluster Overview, Activities, and Available Info"	Michelle Guy, USGS
October 2, 2018	"Develop Intelligence Overview—Managed Learning Solutions"	Sarah Battani, Develop Intelligence

Table 1.4b. Software Development and Information Technology Operations (DevOps) System Administration (SysAd) and Developer Sync topics for fiscal year 2018.

[USGS, U.S. Geological Survey; CHS, Cloud Hosting Solutions; WAF, web application firewall; NASA, National Aeronautics and Space Administration; AWS, Amazon Web Services]

Date	Title	Speaker
October 3, 2017	"Automating Esri Services with Jenkins"	Robert Djurasaj, USGS
November 7, 2017	"Demo and Discussion on GIT Hosting and Version Control"	George Rolston, USGS
December 5, 2017	"CHS Cloudfront/WAF Service"	Jonathan Russo, USGS
February 6, 2018	"Git"	Eric Martinez, USGS
	"Demonstration of QASymphony"	Soraly Mercedes, QASymphony
April 3, 2018	"Tasktop's Presentation Around the Integration Between BMC Remedy and Atlassian JIRA"	Regina Kassar and Mara Puisite, Tasktop
June 5, 2018	"Git" "Supporting NASA's Earth Observing System Data and Information System (EOSDIS)"	Eric Martinez, USGS Dan Pilone, Element84
August 7, 2018	"California Polytechnic State University AWS Migration & Activities"	Paul Jurasin, Theresa May, and Ben Butler, California Polytechnic State University

Table 1.5. Bioinformatics Community of Practice meetings and presentations for fiscal year 2018.

[USGS, U.S. Geological Survey; AWS, Amazon Web Services; SNP, single nucleotide polymorphism]

Date	Title	Speaker
October 17, 2017	"Ongoing Efforts Led by the Alaska Science Center to Provide Templates and Guidance for Genetics Data Release"	Bobbi Pierson, USGS
February 20, 2018	"Christina Kellogg Microbiome Seminar & AWS Demo & List Future Topics"	Christina Kellogg, USGS
April 17, 2018	"Introduction to iMicrobe Platform"	Bonnie Hurwitz, University of Arizona
June 18, 2018	"SNP Topics and SNP Survey"	Scott Cornman, USGS

Table 1.6. Metadata Reviewers Community of Practice meetings and presentations for fiscal year 2018.

[USGS, U.S. Geological Survey; ADIwg, Alaska Data Integration Working Group]

Date	Title	Speaker
October 2, 2017	"Metadata Training"	Group discussion
November 6, 2017	"Discussion of the Biological Data Profile	Pai Yu, Erika Sanchez-Chopitea, and Robin Tillitt, USGS
December 4, 2017	"Data Quality Information"	Group discussion
January 9, 2018	"ISO Metadata in the USGS"	Group discussion
February 5, 2018	"Demo of the New ADIwg Metadata Editor"	Dennis Walworth, USGS, and Josh Bradley, U.S. Fish and Wildlife Service
March 5, 2018	"News and Updates on Metadata Review in the USGS"	Group discussion
April 2, 2018	"USGS Genetics Metadata Workshop Group Resources"	Group discussion
May 7, 2018	"Genetics Guide to Data Release and Associated Data Dictionary"	Barbara Pierson, USGS
June 4, 2018	"Department of the Interior Metadata Implementation Guide"	Ray Obuch, USGS
	"FAIR (Findable, Accessible, Interoperable, Reusable) Metrics"	Group discussion
July 2, 2018	"USGS Web Re-Engineering Team Proposed Metadata Requirements for Legacy Data Sets"	Lisa Zolly, USGS, and group discussion

Table 1.7. Software Development Cluster meetings and presentations for fiscal year 2018.

[USGS, U.S. Geological Survey; HPC/HTC, high performance computing/high-throughput computing]

Date	Title	Speaker
November 2, 2017	"Gitlab Discussion"	Eric Martinez, USGS
November 30, 2017	"Discussion of Goals and Future Topics for the Software Development Cluster"	Group discussion
January 28, 2017	"Summary of USGS HPC/HTC Workshop, USGS EDGE (Equipment and Development Grade Evaluation) Program"	Michelle Guy, USGS
February 22, 2018	"508 Compliance for Websites/Web Applications: Challenges and Approaches"	Rob Miller, USGS
March 29, 2018	"USGS EDGE (Equipment Development Grade Evaluation): What Is It, How Does It Apply to You, and Why You May be Interested in Participating"	Chris Johnson, USGS
April 26, 2018	"Improving the Development Experience on Windows Using HyperV and the Windows Subsystem for Linux"	Scott Lewein, USGS
May 31, 2018	"Software Licensing Aspects"	Leon Foks, USGS
July 26, 2018	"Git Primer, Including Git Fork and Feature Branch Workflow	Carl Schroedl, USGS
August 30, 2018	"Git, a Deeper Dive"	George Rolston, USGS

Table 1.8. Technology Stack Working Group meetings and presentations for fiscal year 2018.

[NetCDF, Network Common Data Form; NDF, XDEDE, Extreme Science and Engineering Discovery Environment; CF, Climate and Forecast]

Date	Title	Speaker
October 12, 2017	“Research Workspace: A Web-Based Tool for Data Sharing, Documentation, Analysis, and Publication”	Rob Bochenek, Axiom Data Science
November 9, 2017	“Jupyter Widgets”	Jason Grout, Bloomberg
December 14, 2017	“Mini-Hack-Session: Developing and Extending Jupyter Notebooks”	Jason Grout, Bloomberg
January 11, 2018	“The Pangeo Project”	Ryan Abernathy, Columbia University; and Matthew Rocklin, Anaconda
February 8, 2018	“The National Data Service Labs Workbench”	Craig Willis, National Center for Supercomputing Applications
March 8, 2018	“Zarr: A Simple, Open, Scalable Solution for Big NetCDF/HDF Data on the Cloud.”	Alistair Miles, University of Oxford
April 12, 2018	“Jetstream: A Free National Science and Engineering Cloud Environment on XSEDE”	Jeremy Fischer, Indiana University
May 10, 2018	“NetCDF-CF Advances—Simple Geometries, Swaths, and Groups”	Dave Blodgett, U.S. Geological Survey; Tim Whiteaker, University of Texas at Austin; Aleksander Jelanek, The Hierarchical Data Format (HDF) Group; and Daniel Lee, EUMETSAT, European Organisation for the Exploitation of Meteorological Satellites
June 14, 2018	“Analysis of Massive Underwater Video Data in the Cloud Using Pangeo”	Tim Crone, Lamont Doherty Earth Observatory
August 9, 2018	“EarthSim: Flexible Environmental Situational Workflows Entirely Within Jupyter Notebooks”	Dharhas Pothina, U.S. Army Engineer Research and Development Center

Table 1.9. Presentations by Community for Data Integration-funded projects at the Earth Science Information Partners Summer Meeting in 2018.

[USGS, U.S. Geological Survey; ISO, International Organization for Standardization]

Title	Speaker
“Community for Data Integration Risk Map Project”	Jeanne Jones, USGS
“An Interactive Web-Based Tool for Anticipating Long-Term Drought Risk”	Caitlin Andrews, USGS
“Integrating Disparate Spatial Datasets from Local to National Scale for Open-Access Web-Based Visualization and Analysis: A Case Study Compiling U.S. Landslide Inventories”	Eric Jones, USGS
“Knowledge Extraction Algorithms (KEA): Turning Literature Into Data”	Daniel Wieferich, USGS
“Content Specifications to Enable USGS Transition to ISO Metadata Standard”	Dennis Walworth, USGS
“Workflows to Support Integrated Predictive Science Capacity: Forecasting Invasive Species for Natural Resource Planning and Risk Assessment”	Kathy Gerst, USA National Phenology Network

Hsu and Colasuonno—**Community for Data Integration 2018 Annual Report**—Open-File Report 2019–1123

ISSN 2331-1258 (online)
https://doi.org/10.3133/ofr20191123