USGS
science for a changing world

Science Synthesis, Analysis, and Research Program

Community for Data Integration 2019 Annual Report

COMMUNITY for DATA INTEGRATION

Open-File Report 2021–1016

Community for Data Integration 2019 Annual Report

By Leslie Hsu and Amanda N. Liford

Science Synthesis, Analysis, and Research Program

Open-File Report 2021–1016

U.S. Geological Survey, Reston, Virginia: 2021

Contents

Figures

Tables

Abbreviations

AI/ML	Artificial Intelligence/Machine Learning
CDI	Community for Data Integration
DevOps	Software Development and Information Technology Operations
eDNA	Environmental DNA
FAIR	Findable, Accessible, Interoperable, Reusable
FY	fiscal year
USGS	U. S. Geological Survey

Community for Data Integration 2019 Annual Report

By Leslie Hsu and Amanda N. Liford

Abstract

The Community for Data Integration is a community of practice whose purpose is to advance the U.S. Geological Survey's data integration capabilities. In fiscal year 2019, the Community for Data Integration held 9 monthly forums, facilitated 11 collaboration areas, held several workshops and training events, and funded 14 projects. The activities supported the U.S. Geological Survey priorities of enabling integrated predictive science, producing FAIR (Findable, Accessible, Interoperable, Reusable) data, building modular and reusable tools, building authoritative national datasets for hazards or assets, and developing tools and methods for biosurveillance of emerging invasive species and health threats. Through these efforts, community members were informed of new and emerging technologies and data topics that helped them in their professional responsibilities.

Introduction

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

The CDI is funded and led by the USGS Science Analytics and Synthesis Program, but membership is voluntary and open to anyone willing to contribute to the community. The CDI has members from the government, as well as from academic, nonprofit, and commercial sectors. In fiscal year (FY) 2019, the CDI welcomed 290 new members to its community of learning and sharing, bringing the total membership to 1,355. Members' self-reported job titles include research scientist, data manager, information technology professional, program manager, software developer, data scientist, and communication specialist, among others.

The CDI acts as a connection point between efforts lead by grassroots practitioners and the priorities of USGS executive leadership. Although CDI activities and topics are driven by members' needs, yearly themes for the proposal process are chosen by CDI's executive sponsors to align with current USGS science goals. The FY 2019 request for proposals themes included (1) FAIR (Findable, Accessible, Interoperable, Reusable) data and tools and (2) methods for biosurveillance of emerging invasive species and health threats. In addition, the theme of the 2019 CDI workshop was "From Big Data to Smart Data," addressing the USGS priority of increasing capabilities for actionable intelligence and 21st-century science.

The following sections explain in more detail the events, outputs, and impact of the CDI community of practice during FY 2019.

Monthly Forums

The CDI holds monthly virtual forums to inform members about new tools, best practices, and relevant standards and policies within the Earth and biological sciences community. The monthly forums serve to reduce barriers to interdisciplinary research and create opportunities for collaborative learning and advancement of Earth and biological science. Topics from FY 2019 included approaches to storing and managing large volumes of scientific data (fig. 1), 3D Elevation Program light detection and ranging (lidar) products and elevation services (fig. 2), the National Aeronautics and Space Administration's Earth Science data collections, and the USGS Water Mission Area's prediction and data integration efforts (fig. 3). Many of the monthly forums included presentations of the outputs from CDI projects funded during FY 2018. Table 1 presents a comprehensive list

of the presentation dates, titles and (or) topics, and speakers. The monthly forums also provide the CDI executive sponsors and coordinators with the opportunity to announce upcoming CDI activities and interact directly with the community. Participation in CDI monthly forums increased by 15 percent from FY 2018 to an average of 112 online participants per virtual meeting.
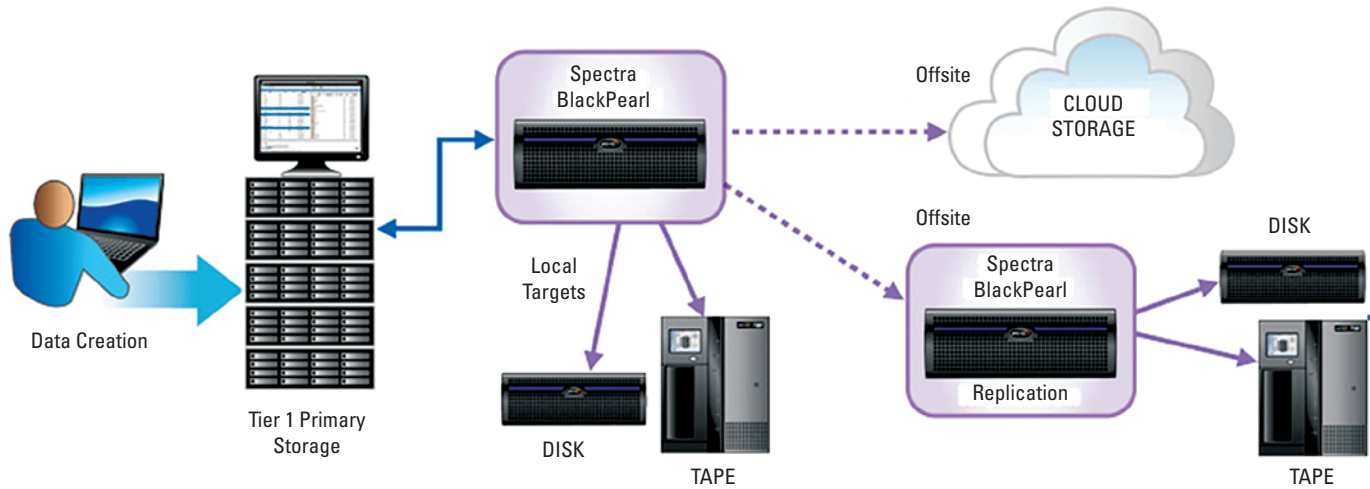
Figure 1.   Image from the October 2018 Monthly Forum presentation "A Cost-Effective Approach to Scientific Data Storage and Management: BlackPearl and Globus" (Matt Davis, U.S. Geological Survey).
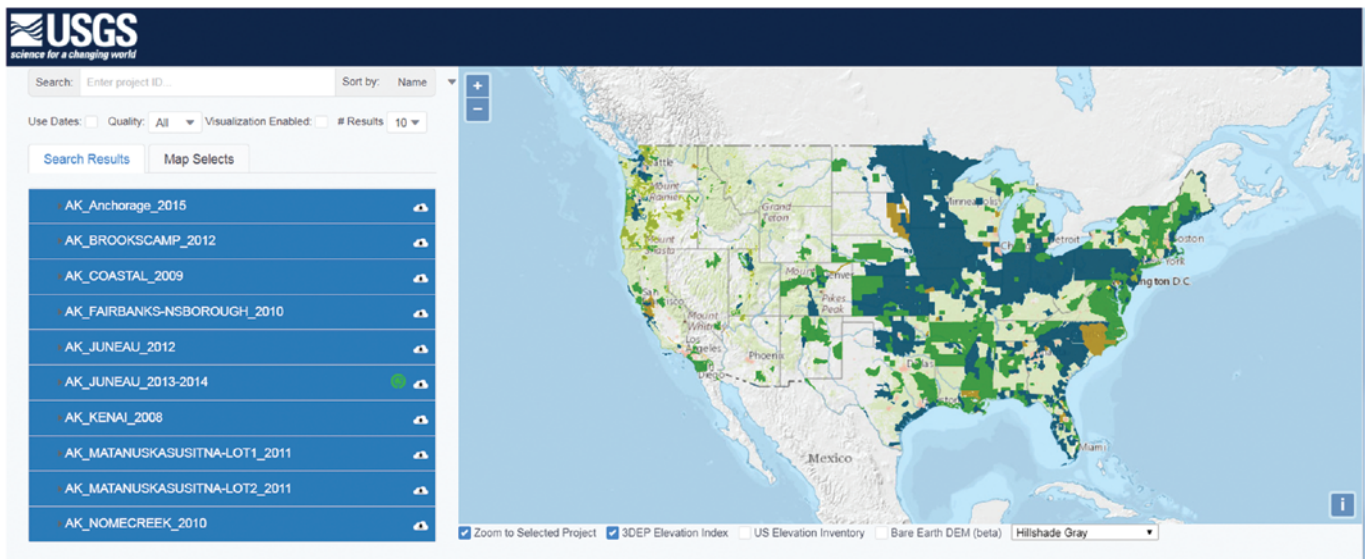
Figure 2.   Screen capture of 3D Elevation Program (3DEP) Lidar Explorer that will enable visualization of all 3DEP lidar data in 2019 from the November 2018 presentation "The New 3D Elevation Program Lidar Products and Elevation Services from the National Map" (Rob Dollison, U.S. Geological Survey).

Figure 3.   Diagram showing integrated priorities from water stakeholders and the need for actionable water intelligence through integrated water analyses, predictions, and data (hydrologic, economic, demographic, environmental, and political) from the September 2019 presentation "USGS Water Prediction Work Program (2WP)" (Katie Skalak, U.S. Geological Survey).

Table 1. Monthly Community for Data Integration (CDI) forum presentations from fiscal year 2019.

[USGS, U.S. Geological Survey; lidar, light detection and ranging; ESIP, Earth Science Information Partners; CUAHSI, The Consortium of Universities for the Advancement of Hydrologic Science; ISO, International Organization for Standardization; NASA, National Aeronautics and Space Administration]

Date	Title(s)	Speaker(s)	Number of attendees
October 10, 2018	"Assembling a National Scale Map of Landslide Inventories from Incomplete and Disparate Spatial Data"	Benjamin Mirus, USGS	113
	"A Cost-Effective Approach to Scientific Data Storage and Management: BlackPearl and Globus"	Matt Davis, Jeff Falgout, Drew Ignizio, USGS	
	"Building a SpatioTemporal Feature Registry (preview)"	Sky Bristol, USGS	
November 14, 2018	"Announcement of the new Artificial Intelligence/Machine Learning Collaboration Area"	Jim Reilly, Director, USGS, and JC Nelson, USGS	112
	"The New 3D Elevation Program Lidar Products and Elevation Services from the National Map"	Rob Dollison, USGS	
	"ESIP Lab Opportunities"	Annie Burgess, ESIP	
February 13, 2019	"CUAHSI Tools for Water Research and Data Publication"	Tony Castronova, CUAHSI	113
	"Workflows to Support Integrated Predictive Science Capacity: Forecasting Invasive Species for Natural Resource Planning and Risk Assessment"	Jake Weltzin, USGS	
	CDI 2019 Workshop, June 4–7, Boulder, Colorado: How to propose topics, submit a demo, request events, sign up, and more	Leslie Hsu, USGS	
March 13, 2019	"Development of the Nonindigenous Aquatic Species Alert Risk Mapper (ARM)"	Wesley Daniel, USGS	114
	"Content Specifications to Enable USGS Transition to ISO Metadata Standard"	Dennis Walworth and Fran Lightsom, USGS	
	"Department of the Interior Risk and the Community for Data Integration Risk Map"	Nate Wood and Jeanne Jones, USGS	
April 10, 2019	"Developing an Interactive Web-based Tool for Anticipating Long-term Drought Risk"	Caitlin Andrews, USGS	120
	"Knowledge Extraction Algorithms (KEA): Turning Literature into Data"	Matthew Neilson, USGS	
	"Mapping Land-use, Hazard Vulnerability and Habitat Suitability Using Deep Neural Networks"	Jon Warrick, USGS	
May 8, 2019	"ICE! Ice Jam Hazard Mobile-Friendly Website"	Hans Vraga and Kathy Chase, USGS	121
	"Investigation of Lidar Data Processing and Analysis in the Cloud"	Jessica Walker, USGS	
	"OpenTopography: Long Tail Topographic Data Management with the Community Dataspace"	Chris Crosby, UNAVCO	
July 10, 2019	"Scientist's Challenge: Data Stories Using Jupyter Notebooks to Increase Reusability of Datasets"	Richie Erickson, USGS	104
	"An Open Online Map of Landslide Occurrence across the U.S. Assembled from Incomplete and Disparate Spatial Data Sets"	Benjamin Mirus, USGS	
	"U.S. Geological Survey National Digital Trails Network"	Greg Matthews and Elizabeth McCartney, USGS	
August 14, 2019	"Exploring NASA's Earth Science Data Collections at earthdata.nasa.gov"	Cynthia Hall, NASA	100
	"Community for Software & Data Integration"	Michelle Guy, Jeremy Newson, and Cassandra Ladino, USGS	
September 11, 2019	"USGS Water Prediction Work Program (2WP)"	Katie Skalak and David Lesmes, USGS	113
	"National Hydrologic Geospatial Fabric: A Framework for the Integration of Water Information"	Roland Viger, USGS	

Collaboration Areas

The CDI is organized into groups—also known as 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 advance data integration in the Earth and biological sciences. Collaboration area membership is voluntary and open to anyone interested in participating. In FY 2019, three new collaboration areas were formed or started interacting with the CDI: Artificial Intelligence/Machine Learning (AI/ML), Fire Science, and Risk Research and Applications. In addition, work continued in the existing collaboration areas of Citizen-Centered Innovation, Data Management, DevOps (Software Development and Information Technology Operations), Environmental DNA (eDNA), Metadata Reviewers, Semantic Web, Software Development, and Technology Stack.

Collaboration area discussions may originate in their separate groups but are often widely applicable to a larger audience. Some recent examples include updates regarding the USGS Publications Warehouse, the purpose and use of the USGS Digital Object Identifier tool, and services provided by the USGS Web Informatics and Mapping Program. Other discussion topics included how to engage social scientists in collaborative research, lessons in stakeholder engagement, and migrating projects from external Git repositories to one of the official USGS source code archives (https://code.usgs.gov). As much as possible, collaboration area activities are documented in the CDI wiki for use by a wider audience (https://my.usgs.gov/confluence/x/yhv1I). The following sections of this report provide a brief description of each collaboration area (table 2) and its activities in FY 2019.

Artificial Intelligence/Machine Learning

The AI/ML collaboration area was created to facilitate sharing and discovery of AI/ML work being done at the USGS, offer learning opportunities, identify AI/ML-related needs, and create connections between practitioners. The AI/ML group began meeting in December 2018 and started by gathering information about AI/ML projects throughout the USGS and sharing AI/ML-related news—such as funding and training opportunities supported by the USGS National Innovation Center (https://www.usgs.gov/centers/innovation)—with the collaboration area itself. Monthly presentations informed members about AI/ML-relevant hardware and software as well as the role of AI/ML within strategic science planning efforts at the USGS and featured various existing USGS and partner projects using AI/ML techniques (fig. 4; app. 1, table 1.1).

Table 2.   Active Community for Data Integration collaboration areas during fiscal year 2019 and contact information for group leads.

[DevOps, Software Development and Information Technology Operations; eDNA, environmental DNA]

| Collaboration area	Group contact(s)
Artificial Intelligence/Machine Learning	Pete Doucette—pdoucette@usgs.gov JC Nelson—jcnelson@usgs.gov
Citizen-Centered Innovation	Sophia Liu—sophialiu@usgs.gov
Data Management	Viv Hutchison—vhutchison@usgs.gov Cassandra Ladino—ccladino@usgs.gov
DevOps	David Hughes—drhughes@usgs.gov Derek Masaki—dmasaki@usgs.gov
eDNA	Scott Cornman—rcornman@usgs.gov
Fire Science	Paul Steblein—psteblein@usgs.gov Mark Miller—mpmiller@usgs.gov
Metadata Reviewers	Fran Lightsom—flightsom@usgs.gov
Risk Research and Applications	Kris Ludwig—kaludwig@usgs.gov David Ramsey—dramsey@usgs.gov
Semantic Web	Fran Lightsom—flightsom@usgs.gov
Software Development	Michelle Guy—mguy@usgs.gov Cassandra Ladino—ccladino@usgs.gov Jeremy Newson—jknewson@usgs.gov
Technology Stack	Richard Signell—rsignell@usgs.gov
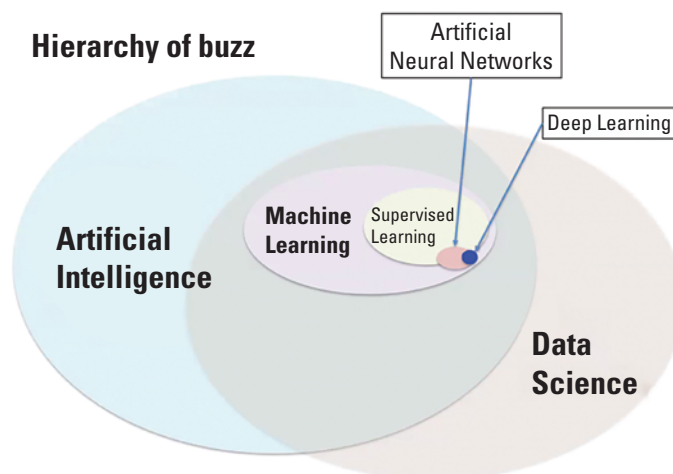
Figure 4. Illustration showing the relationship of different terms to artificial intelligence and data science (Pete Doucette, U.S. Geological Survey).

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." (Hsu and Colasuonno, 2019, p.7) In 2019, the meeting discussions included sharing resources for implementing citizen-centered innovation techniques and information about emerging open innovation projects (app. 1, table 1.2). Other topics included the development of the USGS Open Innovation Strategy, the Office of Science Technology and Policy "Report to Congress for the Crowdsourcing and Citizen Science Act" (Office of Science and Technology Policy, 2019), and a discussion regarding a Department of the Interior Generic Information Collection Request (https://www.doi.gov/generic-and-fast-track-clearances).

Data Management

The Data Management Working Group fosters best practices and collaborative approaches for incorporating data management into USGS science workflows and educating scientists about the value of such methods in research or operational workflows. The group seeks to elevate the practice of data management throughout the USGS Data Lifecycle (Faundeen and others, 2013) because data management is a critical part of the pursuit of science. In FY 2019, the Data Management Working Group continued to keep its members up to date on new tools and methods as well as on data policy and procedures in the USGS. The group hosted presentations on topics including "USGS National Hydrography Data Management," "Becoming a USGS Trusted Digital Repository," "Data Management Staffing in USGS Centers/Programs," and "A First-timer's Experience Publishing Data Processing Software" (app. 1, table 1.3).

DevOps

"The purpose of the DevOps Working Group is to share new techniques and lessons learned using relevant tools and methods."(Hsu and Colasuonno, 2019, p. 8) The DevOps group aims to improve the development-to-deployment efficiency of USGS applications by unifying software development and information technology 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" (Hsu and Colasuonno, 2019, p. 8). In FY 2019, the group hosted presentations on topics such as the "GitLab Web-based DevOps Lifecycle Tool" and "18F and General Services Administration DevOps Best Practices" and held virtual demonstrations of DevOps environments and web portal capabilities from other government agencies such as the National Park Service (app. 1, table 1.4).

eDNA

The purpose of the eDNA Community of Practice is to improve communication, share knowledge, and act as a catalyst for collaboration among people interested in the application of eDNA techniques in Earth-science fields. eDNA is DNA released from an organism into the environment that can be used for the detection and monitoring of native and invasive species (Pilliod and others, 2013). In FY 2019, the group's first year as a CDI collaboration area, presenters shared information on topics including a backpack-style eDNA acquisition device, White Shark eDNA (fig. 5), and choosing the right eDNA assay (app. 1, table 1.5).

Fire Science

The goal of the Fire Science Community of Practice is to share information related to all aspects of fire research, including management and mitigation strategies, ecological effects, socio-economic implications, and advances in detection and monitoring tools. An already established group had been meeting and developing a USGS fire science strategy, but in FY 2019, they joined the CDI in order to reach a larger audience and leverage collaborative infrastructure to share materials and information with participants. Monthly meetings include an update on fire activity in the United States and other announcements related to fire science in addition to a relevant science talk. Science presentation topics included the invasive grass-fire cycle in the southwestern United States, the LANDFIRE remapping effort, means of obtaining access to classified imagery for fire science, and decisions about wildfire risk mitigation (app. 1, table 1.6).

Metadata Reviewers

The purpose of the Metadata Reviewers Community of Practice is to provide a forum for people who review metadata to confer and to establish and maintain consistent metadata review standards throughout the USGS. This group also enables people new to the metadata reviewer role to learn from experienced reviewers. In FY 2019, the group met monthly to discuss various topics including content for a "Reviewing Metadata" section of the USGS Data Management Website, metadata documentation and sharing for long-term monitoring (fig. 6), and checklists for data and metadata review (app. 1, table 1.7).
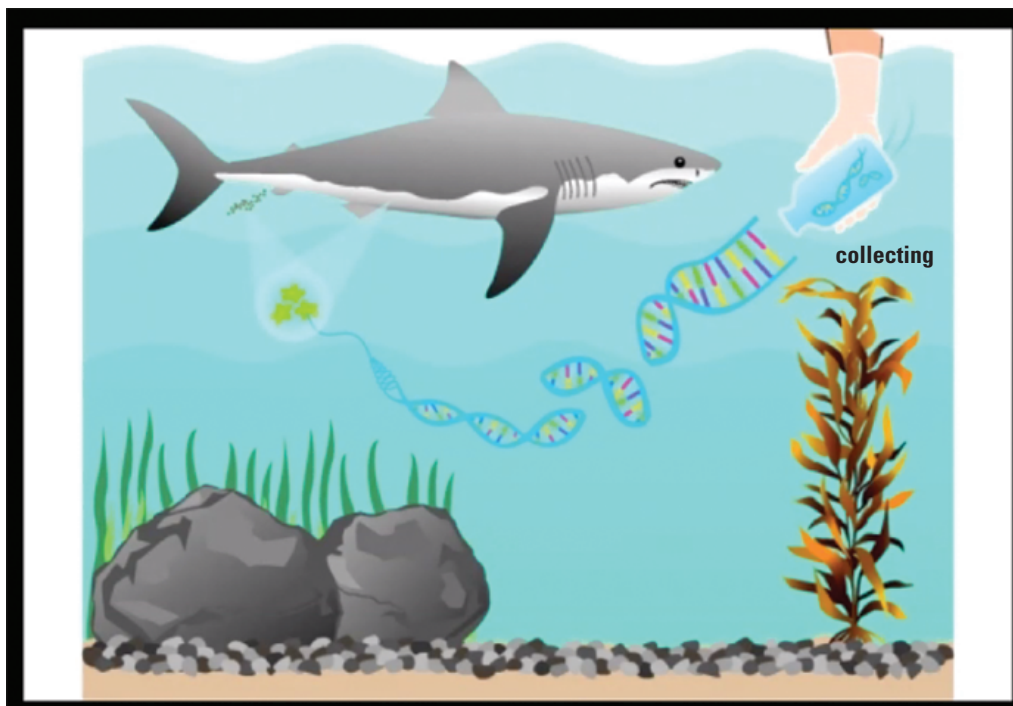

Figure 5.   Image depicting collection of environmental DNA (eDNA) in the ocean (Kevin Lafferty, U.S. Geological Survey).
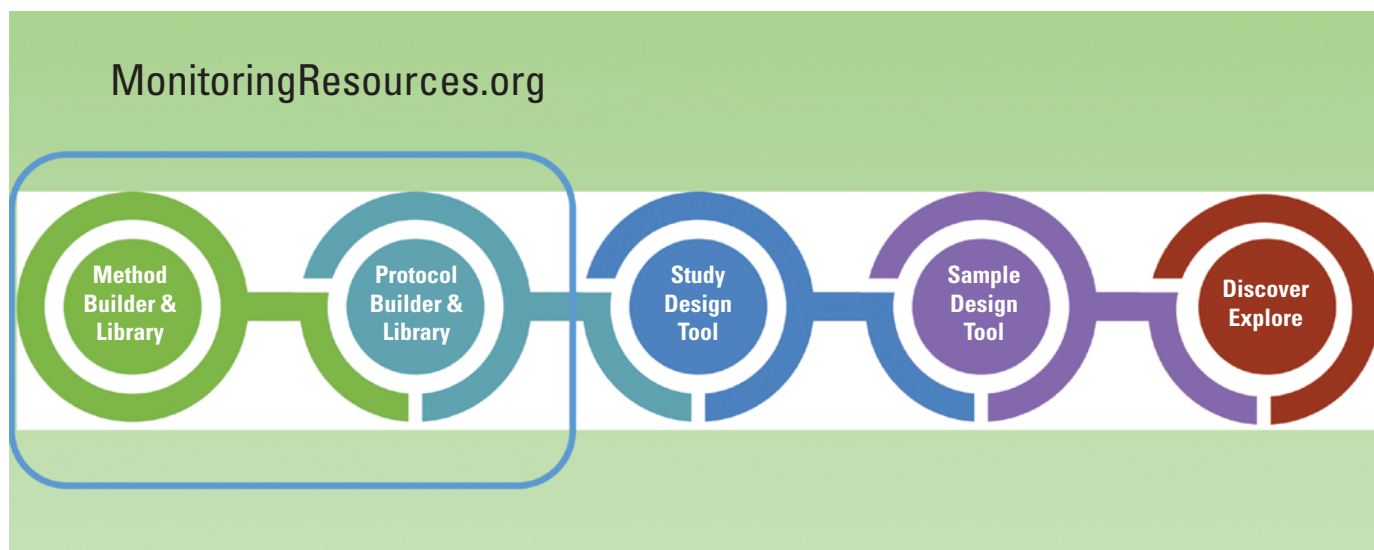
Figure 6. MonitoringResources.org provides a range of tools and services including tools for designing field data collection methods and protocols. Image from the April 2019 presentation titled "MonitoringResources.org: Metadata Documentation and Sharing for Long-Term Monitoring" (Sheryn Olson, U.S. Geological Survey).

Risk Research and Applications

The purpose of the Risk Research and Applications Community is to convene scientists and staff from across the USGS to discuss issues relevant to risk research and applications. Here, risk is defined as "the potential for consequences where something of value is at stake and where the outcome is uncertain, recognizing the diversity of values" (Ludwig and others, 2018, p. 51). The establishment of the community in 2019 was part of the implementation of the 2018 USGS Risk Plan (USGS Science for a Risky World: A USGS Plan for Risk Research and Applications, Ludwig and others, 2018). Monthly meetings include a mix of presentations, discussions, and announcements about relevant resources, events, and activities. Topics explored by the community in FY 2019 included stakeholder engagement, social science, strategic risk research and applications to support U.S. Department of the Interior planning, SAFRR (Science Application for Risk Reduction) projects and scenarios for risk reduction, and novel ways of displaying risk-related data (app. 1, table 1.8). Information from the 2019 Risk Request for Proposals and selected presentations from the Risk annual face-to-face meeting in July 2019 were shared with the wider CDI community.

Semantic Web

"The Semantic Web Working Group is a group of data practitioners working together to explore semantic web technologies to improve the discovery, access, use, and integration of USGS data" (Hsu and Colasuonno, 2019, p. 10). The semantic web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. In FY 2019, discussion topics included semantic approaches to enable USGS data to be FAIR and semantic web elements at the 2019 CDI workshop (app. 1, table 1.9).

Software Development

The Software Development Cluster provides a space for USGS software developers and other interested parties to discuss software release protocols and policies, best practices, software metadata, and software libraries, packages, and tools. Topics in FY 2019 included the USGS code archive (https://code.usgs.gov), big data in cloud computing resources, containers such as Docker (https://www.docker.com/), and desktop installers (app. 1, table 1.10). The group also presented an overview of their activities and the importance of the software component in data integration during the CDI monthly meeting in August.

Technology Stack

"The goal of the Technology Stack Working Group is to explore and share technologies that aid in data discovery, access, and interoperability. The group informs USGS providers and users about tools and techniques to improve efficiency when working with scientific data" (Hsu and Colasuonno, 2019, p. 12).

In FY 2019, the Technology Stack Working Group continued its partnership with Earth Science Information Partners for the Tech Dive webinar series (https://wiki.esipfed.org/Interoperability_and_Technology/Tech_Dive_Webinar_Series). One focus of FY 2019 was science gateways, which are online sites that allow science communities to access shared data, software, computing services, instruments, educational materials, and other resources specific to their disciplines. Science gateway topics covered included JupyterLab extensions, Pangeo for cloud native processing of Earth observation satellite data, and spatiotemporal asset catalog specifications. The other major focus for the group was open knowledge networks and the intersection of open environmental resources and services and structured data for search engine indexing (app. 1, table 1.11).

Special Events and Training

Introduction to Python for Data Science and Matplotlib–Plotting in Python

Many CDI members have requested continuing education on tools for data science and data visualization. After a well-attended, online, self-paced tutorial during FY 2018, we completed two additional online group-learning activities with the same format in FY 2019: "Introduction to Python for Data Science," and "Matplotlib–Plotting in Python." The first course covered Python basics, Python lists, functions and packages, and NumPy. The second activity covered several resources on Matplotlib, a library for creating static, animated, and interactive visualizations in Python. Both activities had approximately 30 participants and used group accountability and discussion on the wiki.

Special Webinar–Connecting Published Data and Information to Scientific Workflows

The CDI also organizes special webinars outside of recurring venues to bring tools and information from external partners to our members. In May 2019, Shanan Peters of the University of Wisconsin presented a talk titled "Connecting Published Data and Information to Scientific Workflows." Peters shared the work that his research group is doing towards establishing an automated system for constructing knowledge bases from scientific publications and using artificial-intelligence-based approaches to text, table, figure, and equation recognition on an enormous database of scientific publications. This research was leveraged by a CDI-supported collaborative project titled "Knowledge Extraction Algorithms (KEA): Turning Literature into Data" (Neilson and others, 2018).

2019 Community for Data Integration Workshop—From Big Data to Smart Data

Biennial in-person CDI workshops bring the community together to discuss current topics and shared challenges, and to share information regarding USGS science priorities. The 2019 workshop was held from June 3–7 at Center Green in Boulder, Colorado. The theme of the workshop was "From Big Data to Smart Data," and the agenda was driven by the needs and interests of the community, with topics highlighting current resources and technologies that could help attendees in their daily work (Hsu, 2021). More than 250 people attended the workshop (235 in person and several dozen online). Attendees included scientists, software developers, data and resource managers, program managers, and others. Themes for breakout sessions included enabling integrated science, computing in the cloud, advancing data management, releasing and preserving science outputs, and improving usability and communication of products.

Outcomes of the workshop included documentation of key take-aways from the topical sessions and input regarding the future direction of the CDI. During sessions, new areas of collaboration and learning were identified, such as data visualization, usability of applications and sites, the importance of software in scientific research, collaborations with external partners, stakeholder engagement, and AI/ML methods. The CDI will build from the results of the workshop to guide future topics, events, and funding opportunities in order to continue to support an integrated science capacity for the USGS.

Making USGS Data FAIR Workshop

The Making USGS Data FAIR Workshop was convened to design a roadmap for implementation of FAIR practices in USGS. The workshop was held September 9–11, 2019, in Fort Collins, Colo. "FAIR is an international set of principles that describes data, other digital products, and methods to promote data discovery and re-use. Applying the FAIR Data principles will improve the application of USGS data, tools, workflows, and systems; creating FAIR data will facilitate integration of USGS science by enhancing data accessibility, discovery, and (re)use" (https://www.usgs.gov/media/images/2019-making-usgs-data-fair-workshop-participants). This workshop was leveraged by a CDI-supported collaborative project titled "Building a Roadmap for Making Data FAIR in the U.S. Geological Survey" (Lightsom and others, 2019).

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." (Hsu and Colasuonno, 2019, p. 15) Project proposals are evaluated based on (1) how well they align with the CDI Science Support Framework (USGS, 2014); (2) how well they satisfy the evaluation criteria of scope, technical approach, project experience and collaboration, sustainability, budget justification, and timeline; and (3) how thoroughly they support the following CDI guiding principles:

- "focus on targeted efforts that yield near-term benefits to Earth and biological science;

- leverage existing capabilities and data;

- implement and demonstrate innovative solutions (e.g., 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." (USGS, 2018, p. 5)

In FY 2019, the CDI continued to support projects that "…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." (USGS, 2017) These focus topics are outlined in a memo regarding the USGS FY18 bureau priorities that support a USGS integrated predictive science capacity (USGS, 2017). Examples of funded projects 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. Specifically, proposals in FY 2019 were encouraged to address one or more of the following themes:

1. Producing FAIR data and tools for integrated predictive science capacity;

2. Reusing or repurposing modular tools such as those developed by previous CDI projects, including the CDI Risk Map (Wood and others, 2018);

3. Building authoritative national datasets for hazards or assets (integrating data and assessing quality); and

4. Developing tools and methods for biosurveillance of emerging invasive species and health threats.

Table 3 provides a summary of each of the 14 CDI projects funded in FY 2019.

Table 3.   Overview of the Community for Data Integration projects funded in fiscal year 2019 (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.

[FAIR, Findable, Accessible, Interoperable, Reusable]

Title	URL	Lead principal investigator	U.S. Geological Survey program
"Open-Source and Open-Workflow Climate Futures Toolbox for Adaptation Planning"	https://www.sciencebase.gov/catalog/item/5cd1ff88e4b09b8c0b7a59a2	Aparna Bamzai	North Central Climate Adaptation Science Center
"Extending ScienceBase for Disaster Risk Reduction"	https://www.sciencebase.gov/catalog/item/5cd20028e4b09b8c0b7a59a8	Joseph A. Bard	Volcano Science Center
"Transforming Biosurveillance by Standardizing and Serving 40 Years of Wildlife Disease Data"	https://www.sciencebase.gov/catalog/item/5cd200d6e4b09b8c0b7a59ac	David S. Blehert	National Wildlife Health Center
"Integrating Short-Term Climate Forecast into a Restoration Management Support Tool"	https://www.sciencebase.gov/catalog/item/5cd20265e4b09b8c0b7a59af	John B. Bradford	Southwest Biological Science Center
"National Public Screening Tool for Invasive and Non-Native Aquatic Species Data"	https://www.sciencebase.gov/catalog/item/5cd20414e4b09b8c0b7a59b7	Wesley M. Daniel	Wetland and Aquatic Research Center
"High-Resolution, Interagency Biosurveillance of Threatened Surface Waters in the United States"	https://www.sciencebase.gov/catalog/item/5cd2055ee4b09b8c0b7a59ba	Sara L. Eldridge	Wyoming-Montana Water Science Center
"Develop Cloud Computing Capability at Streamgages using Amazon Web Services GreenGrass IoT Framework for Camera Image Velocity Gaging"	https://www.sciencebase.gov/catalog/item/5cd2063ae4b09b8c0b7a59bd	Frank L. Engel	Oklahoma-Texas Water Science Center
"Serving the U.S. Geological Survey's Geochronological Data"	https://www.sciencebase.gov/catalog/item/5cd2e8dfe4b09b8c0b7a5c53	Amy K. Gilmer	Geosciences and Environmental Change Science Center
"Establishing Standards and Integrating Environmental DNA (eDNA) Data into the USGS Nonindigenous Aquatic Species Database"	https://www.sciencebase.gov/catalog/item/5cd2e991e4b09b8c0b7a5c56	Margaret E. Hunter	Wetland and Aquatic Research Center
"Subsidence Susceptibility Map for the Conterminous U.S."	https://www.sciencebase.gov/catalog/item/5cd2ea33e4b09b8c0b7a5c59	Jeanne M. Jones	Western Geographic Science Center
"A Generic Web Application to Visualize and Understand Movements of Tagged Animals"	https://www.sciencebase.gov/catalog/item/5cd30962e4b09b8c0b7a5cad	Benjamin Letcher	Leetown Science Center
"Building a Roadmap for Making Data FAIR in the U.S. Geological Survey"	https://www.sciencebase.gov/catalog/item/5cd30aade4b09b8c0b7a5cbb	Frances L. Lightsom	Woods Hole Coastal and Marine Science Center
"Coupling Hydrologic Models with Data Services in an Interoperable Modeling Framework"	https://www.sciencebase.gov/catalog/item/5cd30b8de4b09b8c0b7a5cc1	Richard R. McDonald	Water Resources Mission Area Integrated Modeling and Prediction Division
"Implementing a Grassland Productivity Forecast Tool for the U.S. Southwest"	https://www.sciencebase.gov/catalog/item/5cd30cc7e4b09b8c0b7a5cc6	Sasha C. Reed	Southwest Biological Science Center

Summary

Through the monthly forums, collaboration areas, workshops, projects—as well as through listening to 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 participants and scope, with new collaboration areas in fiscal year 2019 focused on Artificial Intelligence/Machine Learning, fire science, and risk.

The executive sponsors of the CDI continue to challenge the community to use its diverse expertise to help build a U.S. Geological Survey integrated predictive science capacity. To help reach this goal, CDI's activities in fiscal year 2019 focused on the themes of FAIR (Findable, Accessible, Interoperable, Reusable) data, reusing or repurposing modular tools, building authoritative national datasets for hazards or assets, and creating tools and methods for biosurveillance of emerging invasive species and health threats. As the CDI increases its visibility in the U.S. Geological Survey and beyond, we will continue to facilitate activities supporting data and science integration for the Earth and biological sciences.

Acknowledgments

References Cited

Faundeen, J.L., Burley, T.E., Carlino, J.A., Govoni, D.L., Henkel, H.S., Holl, S.L., Hutchison, V.B., Martín, E., Montgomery, E.T., Ladino, C.C., Tessler, S., and Zolly, L.S., 2013, The United States Geological Survey Science Data Lifecycle Model: U.S. Geological Survey Open-File Report 2013–1265, 4 p., accessed February 7, 2021, at https://doi.org/10.3133/ofr20131265.

Hsu, L., 2021, U.S. Geological Survey Community for Data Integration 2019 Workshop Proceedings—From big data to smart data: U.S. Geological Survey Open-File Report 2020–1132, 48 p., accessed February 7, 2021, at https://doi.org/10.3133/ofr20201132.

Hsu, L., and Colasuonno, L., 2019, Community for Data Integration 2018 annual report: U.S. Geological Survey Open-File Report 2019–1123, 26 p., accessed February 7, 2021, at https://doi.org/10.3133/ofr20191123.

Lightsom, F.L., Hutchison, V.B., Faundeen, J.L., Latysh, N., Debrewer, L.M., Govoni, D.L., Bishop, W., and Stall, S., 2019, Building a roadmap for making data FAIR in the U.S. Geological Survey: U.S. Geological Survey ScienceBase data release, accessed November 2, 2020, at https://www.sciencebase.gov/catalog/item/5cd30aade4b09b8c0b7a5cbb.

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.

Neilson, M.E., Wieferich, D.J., Daniel, W.M., Serna, B.S., and Peters, S., 2018, Knowledge extraction algorithms (KEA)—Turning literature into data: U.S. Geological Survey ScienceBase data release, accessed June 12, 2020, at https://www.sciencebase.gov/catalog/item/5acd2680e4b0e2c2dd155dfd.

Office of Science and Technology Policy, 2019, Implementation of Federal prize and citizen science authority—Fiscal years 2017–18: Office of Science and Technology Policy, accessed November 2, 2020, at https://www.whitehouse.gov/wp-content/uploads/2019/06/Federal-Prize-and-Citizen-Science-Implementation-FY17-18-Report-June-2019.pdf.

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., accessed June 12, 2020, at https://doi.org/10.3133/fs20123146.

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

U.S. Geological Survey, [USGS], 2018, U.S. Geological Survey (USGS) Community for Data Integration (CDI) request for proposals (RFP) for fiscal year 2019: U.S. Geological Survey, 19 p., accessed February 8, 2021, at https://my.usgs.gov/confluence/display/cdi/2019+Proposals?preview=%2F604342343%2F611189036%2FCDI_FY19_Request_for_Proposals_final.pdf.

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

Appendix 1. Presentations and Speakers

This appendix lists presentation topics and speakers from the different collaboration area meetings for fiscal year 2019.

Table 1.1. Artificial Intelligence/Machine Learning meetings and presentations from fiscal year 2019.

[USGS, U.S. Geological Survey; AI, Artificial Intelligence; ML, Machine Learning; USFWS, U.S. Fish and Wildlife Service]

Date	Subject	Speaker(s)
December 11, 2018	Inaugural meeting	Tim Quinn and JC Nelson, USGS
February 12, 2019	"Opportunities at the USGS Innovation Center related to AI/ML"	John Stock, USGS
	"Ruminations on AI and Land Imaging"	Pete Doucette, USGS
March 12, 2019	"Tallgrass, a New Machine for AI at the USGS"	Jeff Falgout, USGS
	"AI for Integrated Environmental Modeling & Forecasting, and Overview of AI for Ecosystem Services"	Ken Bagstad, USGS
May 14, 2019	"XGBoost in Continuous Change Detection and Classification (CCDC)"	Chris Barber, USGS
	"Deep Learning to Quantify Benthic Habitat"	Peter Esselman, USGS
June 11, 2019	Strategic Science Planning at USGS	Pete Doucette, USGS
August 13, 2019	"Continuous Streamflow and Nearshore Wave Monitoring from Time-lapse Cameras Using Deep Neural Networks"	Daniel Buscombe, Northern Arizona University
September 10, 2019	"Utilizing Deep Neural Networks for Landscape Conservation: An Application of Google's Tensorflow for a Cannabis Production Inventory in Northern California"	Daryl Van Dyke, USFWS

Table 1.2. Citizen-Centered Innovation meetings and presentations from fiscal year 2019.

[DOI, U.S. Department of the Interior; USGS, U.S. Geological Survey; OSTP, Office of Science and Technology Policy]

Date	Subject	Speaker
May 15, 2019	"DOI Generic Information Collection Request"	Sophia Liu, USGS
	"USGS Open Innovation Strategy"	Sophia Liu, USGS
April 17, 2019	"OSTP Report to Congress for the Crowdsourcing and Citizen Science Act"	Sophia Liu, USGS
	Continued discussion of "DOI Generic Information Collection Request"	Sophia Liu, USGS

Table 1.3.   Data Management monthly meeting presentations from fiscal year (FY) 2019.

[FAIR, Findable, Accessible, Interoperable, and Reusable; USGS, U.S. Geological Survey; ISO, International Organization for Standardization; ADIwg, Alaska Data Integration Working Group; CDI, Community for data Integration; DMWG, Data Management Working Group; RFP, request for proposals]

Date	Title(s)	Speaker(s)
October 29, 2018	"Trust Repositories, but FAIR-ify"	Wade Bishop, University of Tennessee
December 10, 2018	"Updates about Publications Warehouse"	Kelly Haberstroh, USGS
	"Updates on the ISO for USGS: Content Specifications and Current Status of ADIwg"	Dennis Walworth, USGS
	"Updates to the Digital Identifier Tool"	Lisa Zolly, USGS
March 11, 2019	"Community Discussion on Goals for the CDI Workshop DMWG Session"	Cassandra Ladino, USGS
April 9, 2019	"Becoming a USGS Trusted Digital Repository"	John Faundeen and Natalie Latysh, USGS
	"Update on Progress for USGS Data Manager Position Description Series"	Viv Hutchison and John Faundeen, USGS
May 13, 2019	"Behind the Scenes at ScienceBase: How Data Release Happens in Your USGS Trusted Digital Repository (TDR)"	Tamar Norkin, USGS
July 8, 2019	"Data Manager Position Descriptions: A Brief Overview"	Viv Hutchison, USGS
	"Data Management Staffing in USGS Centers/Programs"	Madison Langseth, USGS
August 12, 2019	"USGS National Hydrography Data Management"	Karen Adkins and Jerry Ornelas, USGS
	"A First-timer's Experience Publishing Data Processing Software"	Emily Sturdivant, USGS
September 9, 2019	"Data Management Training Clearinghouse"	Karl Benedict, USGS
	"USGS-Specific Training Materials and Resources and CDI FY20 RFP Teaser"	Madison Langseth, USGS

Table 1.4.   Software Development and Information Technology Operations (DevOps) topics from fiscal year 2019.

[USGS, U.S. Geological Survey; CI/CD, Continuous Integration and Continuous Delivery; HUD, Department of Housing and Urban Development; GSA, General Services Administration; NPS, National Park Service; CHS, Cloud Hosting Solutions]

Date	Title	Speaker(s)
October 2, 2018	"Develop Intelligence Overview-Managed Learning Solutions"	Sarah Battani, USGS
November 6, 2018	"https://recreation.gov—Portal for Outdoor Recreation"	Shums Hoda, Booz Allen Hamilton
February 5, 2019	"USGS Web Informatics and Mapping Program"	Hans Vraga, USGS
March 5, 2019	"Technical Overview of the DevOps Environment and CI/CD pipeline"	Martin Folkoff, Booz Allen Hamilton
April 2, 2019	"Department of Housing and Urban Development DevOps"	Kevin Portanova and Mel Hurley, HUD
May 7, 2019	"GitLab Web-based DevOps Lifecycle Tool"	Kyle Goodwin, GitLab
July 2, 2019	"18F/GSA DevOps Best Practices"	Peter Burkholder, GSA/18F
August 6, 2019	"National Park Service Vector Tiling Activities"	Jim McAndrew, NPS
September 3, 2019	"USGS Bird Banding Lab ReportBand Application"	George Ralston and other staff from CHS (USGS)

Table 1.5. Environmental DNA (eDNA) Community of Practice meetings and presentations from fiscal year 2019.

[USGS, U.S. Geological Survey; UCSB, University of California, Santa Barbara]

Date	Title	Speaker
November 13, 2018	"Data on a Backpack-style eDNA Acquisition Device"	Austen Thomas, Smith-Root
February 19, 2019	"CALeDNA (California Environmental Data)"	Rachel Meyer, University of California Conservation Genomics Consortium
April 16, 2019	"White Shark eDNA"	Kevin Lafferty, USGS/UCSB
May 21, 2019	"Choosing the Right eDNA Assay: Developing Standards for Limit of Detection and Limit of Quantification"	Chris Merkes, USGS

Table 1.6. Fire Science Community of Practice meetings and presentations from fiscal year 2019.

[USGS, U.S. Geological Survey]

Date	Title(s)	Speaker(s)
February 12, 2019	"Effects of Prescribed Fire on San Francisco Gartersnake Survival and Movement"	Brian Halstead, USGS
March 20, 2019	Updates from the Fire Science Strategic Planning Team	Paul Steblein, USGS
April 9, 2019	"Changing Environmental Drivers, Tipping Points, and Resilience in Fire-prone Ecosystems"	Craig Allen, USGS
May 21, 2019	"Slowing the Invasive Grass-fire Cycle in Dryland Ecosystems of the Southwestern U.S."	Seth Munson, USGS
July 16, 2019	"LANDFIRE Remap—Developing a New Baseline Product Suite"	Joshua Picotte, USGS
	"Arsenic and Old Mines–Wildfire Remobilizes Historical Mining Waste"	Sheila Murphy, USGS
August 20, 2019	"Sunk Cost, Well Spent: Enhancing Interagency Remote Sensing with the Civil Applications Committee"	Dan Opstal, USGS
September 17, 2019	"Decisions About Wildfire Risk Mitigation on Private Property: the Role of Spatial Effects, Risk Perceptions, and Systematic Data Collection"	James Meldrum, USGS

Table 1.7.   Metadata Reviewers meetings and presentations from fiscal year 2019.

[USGS, U.S. Geological Survey; ISO, International Organization for Standardization; FAIR, Findable, Accessible, Interoperable, and Reusable; CDI, Community for Data Integration]

Date	Title or topic(s)	Speaker(s)
October 1, 2018	Current state of metadata for data releases in USGS, status of ISO Content Specs project, and a new effort to enable FAIR principles in USGS	Group discussion
November 5, 2018	Review of checklists for data review and metadata review and discussion of how we review data and metadata in our science centers	Group discussion
December 3, 2018	Using links to publications in the Process Step; follow up from data and metadata review checklists	Group discussion
February 4, 2019	Discussion of guidance on the volume of data necessary to trigger a separate data release and how authors should reference data that is not publicly available when writing a manuscript	Group discussion
March 4, 2019	Discussion of location parameters in metadata and data; discussion of a Jupyter notebook for quick evaluation of large data files	Group discussion
April 1, 2019	"MonitoringResources.org: Metadata Documentation and Sharing for Long-term Monitoring"	Sheryn Olson, USGS
May 6, 2019	"How the USGS Science Data Catalog Could Use the Keywords in Metadata Records to Improve Data Discovery"	VeeAnn Cross and Peter Schweitzer, USGS
July 1, 2019	Wrap-up from CDI breakout session	Group discussion
August 5, 2019	"Proposed 'Reviewing Metadata' Page on the USGS Data Management Website"	Madison Langseth, USGS and group discussion

Table 1.8.   Risk meetings and presentations from fiscal year (FY) 2019.

[USGS, U.S. Geological Survey; SAFRR, Science Application for Risk Reduction; DOI, U.S. Department of the Interior; Bsal, *Batrochochytrium salamandrivorans*]

Date	Title or topic(s)	Speaker(s)
March 21, 2019	Community of Practice Kickoff Meeting	Kris Ludwig, USGS
March 26, 2019	FY19 Request for Proposals Informational Webinar	Kris Ludwig, USGS
April 18, 2019	"Shared Learning and Intentional Collisions—Benefits of a Community of Practice"	Leslie Hsu, USGS
	"Relationships with Key Stakeholders: Examples from ShakeCast"	David Wald, USGS
May 16, 2019	"Social Science 101"	Sara McBride, USGS
	"SAFRR Projects & Scenarios for Risk Reduction"	Dale Cox, USGS
June 20, 2019	"The USGS Earthquake Hazards Program 'Engineering & Risk' Project"	Nico Luco, USGS
	"Strategic Risk Research and Applications to Support DOI Planning"	Nate Wood, USGS
July 18, 2019	"Quantifying Rock Fall Hazard and Risk to Roadways in National Parks: Yosemite National Park Pilot Project"	Brian Collins, USGS
	"Re-visiting Bsal Risk: How 3 Years of Pathogen Surveillance, Research, and Regulatory Action Change our Understanding of Invasion Risk of the Exotic Amphibian Pathogen *Batrochochytrium salamandrivorans*"	Dan Grear, USGS
	"The State of Our Coasts: Coastal Change Hazards Stakeholder Engagement & User Need Assessment"	Emily Himmelstoss, USGS
	"Assessing the Risk of Global Copper Supply Disruption from Earthquakes"	Kishor Jaiswal, USGS
	"Communications of Risk-Uranium in Groundwater in Northeastern Washington State"	Sue Kahle, USGS
	"Cascading Communication: Exploring How Scientific Research Affects Policy and Earthquake Preparedness Communication in Areas of Induced Seismicity"	Sara McBride, USGS
	"Ecological Forecasts for Risk Management"	Alyssa Rosemartin, National Phenology Network
	"Hazard Exposure Analyst Tool (HEAT)"	Jason Sherba, USGS
July 19, 2019	Special Topic Session: Panel discussion on stakeholder engagement	Amanda Cravens, Alice Pennaz, Rudy Schuster, and Anne Wein, USGS
August 15, 2019	"Rigorously Quantifying the Coastal Hazard Risk Reduction Provided by US Coral Reefs"	Curt Storlazzi, USGS
	"I Have an Idea! Alternative Ways of Displaying our Data"	Athena Clark, USGS
September 19, 2019	"The WiRē Team: a Long-term Research-practice Collaboration for Supporting Wildfire Adaptedness in Wildland-Urban Interface (WUI) Communities"	James Meldrum, USGS
	"Community-Centered Climate Planning for People and Parks"	Emily Brooks, USGS

Table 1.9.   Semantic Web meetings and presentations from fiscal year 2019.

[USGS, U.S. Geological Survey; CDI, Community for Data Integration; FAIR, Findable, Accessible, Interoperable, Reusable]

Date	Topic	Speaker(s)
October 11, 2018	Semantic approaches to enable USGS data to be FAIR (Findable, Accessible, Interoperable, Reusable)	Group discussion
February 14, 2019	CDI FAIR in USGS Proposal and Workshop Ideas	Group discussion
March 14, 2019	Semantic Web elements at 2019 CDI Workshop	Group discussion
June 13, 2019	Contributing to the USGS FAIR Roadmap	Group discussion

Table 1.10.   Software Development meetings and presentations from fiscal year (FY) 2019.

[USGS, U.S. Geological Survey; CDI, Community for Data Integration; NGTOC, National Geospatial Technical Operations Center]

Date	Title or topic(s)	Speaker(s)
October 25, 2018	"Migrating Projects from Other Git Repositories to code.usgs.gov"	Eric Martinez, USGS
February 28, 2019	"Cloud and Big Data in the Cloud"	Cassandra Ladino, USGS
March 28, 2019	"Brainstorming and Planning for the 2019 CDI Annual Workshop"	Group discussion
April 25, 2019	"Collaboration through the Software Dev Cluster, Desktop Installers and Script Signing"	Cassandra Ladino, USGS
May 23, 2019	"Docker Basics for Code Development"	Michelle Guy, Jeremy Newson, and Cassandra Ladino, USGS
June 27, 2019	"CDI Workshop Report Out"	Cassandra Ladino, USGS
July 25, 2019	"Deeper Dive into Containers and Docker, Water Mission Area and NGTOC"	Carl Schroedl and Ivan Suftin, USGS
August 22, 2019	CDI FY20 Request for Proposals	Group discussion
September 26, 2019	"Databases and Beyond!"	Cassandra Ladino, USGS

Table 1.11.　Technology Stack meetings and presentations from fiscal year 2019.

[USGS, U.S. Geological Survey; ADIwg, Alaska Data Integration Working Group; USFWS, U.S. Fish and Wildlife Service]

Date	Title	Speaker(s)
October 11, 2018	"SpatioTemporal Feature Registry: ESIP Idea Campaign and Working Example in USGS"	Sky Bristol, USGS
November 8, 2018	"Intake: Lightweight Tools for Loading and Sharing Data in Data Science Projects"	Martin Durant, Anaconda
December 13, 2018	"Developing JupyterLab Extensions"	Ian Rose, University of California, Berkeley
February 14, 2019	"Cloud Native Geoprocessing of Earth Observation Satellite Data with Pangeo"	Scott Henderson, University of Washington
March 14, 2019	"Integrating SciServer and OceanSpy to Enable Easy Access to Oceanographic Model Output"	Mattia Almansi, Johns Hopkins University
April 11, 2019	"Pachyderm"	John Karabaic, Pachyderm
May 9, 2019	"The SpatioTemporal Asset Catalog (STAC) Specification"	Chris Holmes, Planet Labs
June 6, 2019	"Google Dataset Search: Facilitating Data Discovery in an Open Ecosystem"	Chris Gorgolewski, Google
July 11, 2019	"ADIwg Open Source Metadata Toolkit"	Josh Bradley, USFWS and Dennis Walworth, USGS
August 8, 2019	"The Challenge of Location and How Discrete Global Grid Systems Can Enable Spatial Data Integration"	Matthew Purss, Geoscience Australia
September 12, 2019	"STARE: SpatioTemporal Adaptive Resolution Encoding for Scalable Integrative Analysis"	Michael Rilee, Rilee Systems Technologies, LLC