

Science Synthesis, Analysis, and Research Program

U.S. Geological Survey Community for Data Integration 2019 Workshop Proceedings— From Big Data to Smart Data



Open-File Report 2020–1132

U.S. Geological Survey Community for Data Integration 2019 Workshop Proceedings— From Big Data to Smart Data

By Leslie Hsu

Science Synthesis, Analysis, and Research Program

Open-File Report 2020–1132

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: 2021

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.

Although this information product, for the most part, is in the public domain, it also may contain copyrighted materials as noted in the text. Permission to reproduce copyrighted items must be secured from the copyright owner.

Suggested citation:

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., <https://doi.org/10.3133/ofr20201132>.

ISSN 2331-1258 (online)

Contents

Abstract.....	1
Introduction.....	1
Presentations.....	3
Welcome and Opening Remarks—From Big Data to Smart Data	3
Turning Your Data Into Real-Time Actionable Insights.....	3
Lightning Talks—Part 1	3
Partner Presentations	3
Data + Community = Action.....	3
Let’s Play: How the ESIP Lab Is Supporting Innovation and Ingenuity in Earth-Science Technology Development.....	4
Science Gateways Community Institute	4
Components of Integrated Science	5
USGS National Hydrography Infrastructure	5
USGS Risk Map—A Foundation for Integrated-Risk Science	5
Advanced Research Computing (ARC): Providing Essential Tools for Integrated Science	5
USGS Cloud-Hosting Solutions Update.....	5
“Thinking in Systems”: ScienceBase’s Role in Supporting Integrated Science and Linked-Information Management in the USGS.....	5
GeoPlatform	5
FAIR Data Session	6
FAIR 101	6
This Is So UnFAIR!	6
Mineral Resources Program: Delivering Data FAIR and Square.....	6
Enterprise Tools for Documenting Natural Resources Monitoring Efforts	7
Evaluation of Data FAIR-ness From USGS Data Users’ Perspectives	7
Lightning Talks—Part 2.....	7
USGS Big Data to Smart Data.....	7
Ecological Forecasting—Making Smart Data From Big Data.....	7
Big Data From a Big Disaster: UAS (Unmanned Aircraft System) Data Collection, Processing, and Dissemination During the 2018 Kilauea Eruption	7
From Big Data to Smart Data in Remote Sensing.....	8
Process-Guided Deep-Learning Predictions of Lake-Water Temperature.....	8
Evidence-Based Policy and Department of Interior Data Governance.....	8
Topical Sessions	8
Advancing Data Management.....	8
Data-Management Workflow Show and Tell.....	8
Records-Management and Data-Management Connections.....	8
Advanced Approaches to Data Management: Exploring Services and APIs	9
Content Specifications for ISO Metadata Standard	9
Computing in the Cloud.....	9
Cloud Hosting Solutions—Service Offerings in the USGS Cloud	9
Cloud-Hosting Solutions 101: Demystifying the Cloud	9

Let's Talk Cloud.....	10
Let's Talk Cloud—Applications.....	10
Enabling Integrated Science	10
Integrating Data and Models for Next-Generation Predictive Science.....	10
Exploring the Landscape of Scientific-Computing Tools for Large-Scale Analytics, Machine Learning, and Integrated Modeling.....	11
Visualization Tools You Can Use: CDI Risk Map, TerriaMap, Tableau.....	11
Improving Usability and Communication	11
Want to Know How to Make Your Data Applications Delightful? Usability Can Help!	11
SGCI Communication Session—Building your Value Proposition	11
Releasing and Preserving Science Outputs.....	11
Preservation and Digitization of Physical Materials.....	11
Progress on Handling Large Data in the USGS and What's Ahead.....	12
Data and Metadata Review and Creation	12
Software-Release Questions Answered	12
Collaboration Area Meetings.....	12
Citizen-Centered Innovation Working Group	12
Data-Management Working Group	13
Software Development Cluster	14
Birds of a Feather Discussions.....	14
Trainings	15
Introduction to R.....	15
USGS Software-Release Practicum.....	15
ISO Metadata-Content Specification Workshop.....	15
High-Performance Computing.....	16
DataBlast.....	16
USDA (U.S. Department of Agriculture) Forest Service National Riparian Inventory Base Map.....	16
New Imagery Data System for the Coastal/Marine Hazards and Resources Program.....	17
Preservation and Publication of Historical Seismic Data from the U.S. Geological Survey.....	17
Transforming Biosurveillance by Standardizing and Serving 40 Years of Wildlife-Disease Data.....	17
Integrated Modeling: Making Data and Models FAIR With AI and the Semantic Web.....	18
Integrating Short-Term Climate Forecasts Into a Restoration-Management Support Tool	18
Developing Modern Web Pages by Using the U.S. Web Design System.....	18
A Community Supporting a Community: How CDI Members Can Use ESIP Lab Technology Development, Sharing, and Evaluation Services	19
Ice Jams—Mobile-Friendly Website for Collection and Display of Real-Time and Historical Ice-Jam Information.....	19
ScienceBase and EROS Collaboration to Release UAS Imagery.....	19
Have You SEINeD Your Data Lately? A National Public Screening Tool for Invasive and Nonnative Aquatic-Species Data	19
NHDPlus High Resolution	20

High-Resolution Interagency Biosurveillance of Threatened Surface Waters in the United States.....	20
Develop Cloud Computing Capability at Streamgages Using Amazon Web Services Greengrass IoT Framework for Camera Image Velocity Gaging.....	20
USGS Cloud-Hosting Solutions Support 21st-Century Science.....	21
Establishing Standards and Integrating Environmental DNA (eDNA) Data Into the USGS Nonindigenous Aquatic-Species Database.....	21
Serving the U.S. Geological Survey’s Geochronological Data.....	21
Gridding Airborne Radiometric Data Using the Stan Probabilistic Programming Language: An Example of Bayesian Statistical Modeling With RStudio and ArcGIS Pro.....	21
SageDAT: Data and Tools to Support Collaborative Sagebrush-Ecosystem Conservation and Management.....	22
DataAtRisk.org and Its Community-Driven Data-Rescue Nomination Tool.....	22
Measuring Sustainability of Community for Data-Integration Projects.....	22
USGS GGGSC Data Management and Spatial Studies (DMSS), Geophysical Inventory(s), Mobile Data Collection, and Automation.....	23
Subsidence-Susceptibility Map for the Conterminous United States.....	23
Free Chocolate!!! Short Digital Object Identifier Usability Test Required.....	23
A Generic Web Application to Visualize and Understand Movements of Tagged Animals.....	23
Building a Road Map for Making Data FAIR in the U.S. Geological Survey: A 2019 CDI Project.....	24
Improving a Classification System to Support FAIR Coastal and Marine Data.....	24
Coupling Hydrologic Models With Data Services in an Interoperable Modeling Framework.....	24
Making Connections: The U.S. Geological Survey’s National Digital Trails Network.....	25
The Oceanographic Model and Data Portal: One Example of Integrating Interoperable Data.....	25
ScienceBase as a Platform for Data Release: Workflow Automation.....	25
Advanced Computing Cooperative.....	26
Reproducible Data Pipelines for Scientific Analyses.....	26
Implementing a Grassland-Productivity Forecast Tool for the U.S. Southwest.....	26
GIS and Information Management Project—Putting Mineral Resources Program on the Map.....	27
UAS Rad Cal: Open-Source Radiometric-Calibration Software.....	27
Pangeo: A Platform for Big-Data Geoscience on the Cloud.....	27
First Comprehensive Nonnative Species List for the United States, Segregated by Major Regions.....	28
Species-Occurrence Data for the Nation.....	28
Automated ScienceBase Data-Release Workflow: A Script to Update Metadata and Populate SB Pages.....	28
Decoding NHD’s VisibilityFilter Attribute: What Is It? Where Is It Available? and What Is the Accuracy?.....	29
Connecting Data to the National Hydrography Dataset.....	29
ISO Made Easy: Content Specifications to Guide Metadata Authorship.....	29
SHIRA Risk Mapper—Visualizing Multihazard Risk Exposure of DOI Lands, Populations, Assets, and Resources.....	30

Summary of Workshop Outcomes	30
Acknowledgments	30
References Cited.....	31
Appendix 1. Agenda	32
Appendix 2. Attendees.....	34
Appendix 3. Key Take-aways.....	41
Appendix 4. Interactive Questions and Comments.....	46
Appendix 5. Community for Data Integration and Science Support Framework.....	48

Figures

1. Attendees at the 2019 Community for Data Integration Workshop during a break in the plenary schedule. Photograph by Jacob Massey.....	2
2. “Napkin drawings” inspired by the presentation by the Science Gateways Community Institute. Photograph by Leslie Hsu.....	4
3. Shelley Stall addressing the plenary audience about FAIR data. Photograph by Leslie Hsu	6
4. Speed data-ing at the Data Management Working Group breakout session. Pairs of attendees discuss their common interests and challenges in data management. Photograph by Vivian Hutchison	13
5. List of Birds of a Feather discussions during the week. Photograph by Leslie Hsu	14
6. Meeting attendees discuss their work at the DataBlast poster and demonstration session. Photograph by Jacob Massey.....	16
7. The Community for Data Integration Science Support Framework.....	48

Conversion Factors

International System of Units to U.S. Customary Units

	Length	
	Multiply	To obtain
meter (m)	3.281	foot (ft)
kilometer (km)	0.6214	mile (mi)
kilometer (m)	0.5400	mile, nautical (nmi)

Abbreviations

ACC	Advanced Computing Cooperative
AGRS	airborne gamma-ray spectrometry
AI	artificial intelligence
API	Application Programming Interface
app	application
ARC	Advanced Research Computing
ARIES	Artificial Intelligence for Ecosystem Services
AWS	Amazon Web Services
BISON	Biodiversity Information Serving Our Nation
CDI	Community for Data Integration
CF	Climate and Forecast
CHS	Cloud-Hosting Solutions
CLI	Command Line Interface
CMECS	Coastal and Marine Ecological Classification Standard
CMGDS	Coastal and Marine Sciences Data System
CMGP	Coastal and Marine Geology Program
CMHRP	Coastal and Marine Hazards and Resource Program
COP	Community of Practice
CSDMS	Community Surface Dynamics Modeling System
DMSS	Data Management and Spatial Studies
DN	Digital Number
DNA	deoxyribonucleic acid
DOE	Department of Energy
DOI	Department of the Interior
EarthMAP	Earth Monitoring Analyses and Prediction
EarthMRI	Earth Mapping Resources Initiative
eDNA	Environmental DNA
EROS	Earth Resources Observation and Science
ESIP	Earth Science Information Partners
ETL	Extract, Transform, Load
FAIR	findable, accessible, interoperable, and reusable
FGDC	Federal Geographic Data Committee
GBIF	Global Biodiversity Information Facility
GGGSC	Geology, Geophysics, and Geochemistry Science Center

GIS	geographic information system
GeoPlatform	Geospatial Hosting Platform
GUI	Graphical User Interface
HPC	High-Performance Computing
HRD	Hydrography Referenced Data
HRT	Hydrography Referencing Tool
HTC	High-throughput Computing
IoT	Internet of Things
IPDS	Information Product Data System
iRIC	International River Interface Cooperative
ISO	International Organization for Standards
IT	information technology
LIMS	Laboratory Information Management System
MRP	Mineral Resources Program
NAS	Nonindigenous Aquatic Species
NCAR	National Center for Atmospheric Research
NDT	National Digital Trails Network
netCDF	Network Common Data Form
NGDB	National Geochronological Database
NGP	National Geospatial Program
NHD	National Hydrography Dataset
NHDPlus HR	National Hydrography Dataset Plus High Resolution
NHI	National Hydrography Infrastructure
NLCD	National Land Cover Database
NOAA	National Oceanographic and Atmospheric Administration
NPS	National Park Service
NUPO	National UAS Project Office
NWHC	National Wildlife Health Center
OSQI	Office of Science Quality and Integrity
PyMT	Python Modeling Toolkit
RBDM	Riparian Buffer Delineation Model
SB	ScienceBase
SEG-Y	Society of Exploration Geophysicists geophysical data format
SGCI	Science Gateways Community Institute
SHIRA	Strategic Hazard Identification and Risk Assessment

SSF	Science Support Framework
SurfBOARD	Surface Velocity Workgroup
THREDDS	Thematic Real-Time Environmental Distributed Data Services
TIFF	Tagged Image File Format
TRAILS	Trail Routing, Analysis, and Information Linkage System
UAS	Unmanned Aircraft System
USDA	U.S. Department of Agriculture
USGS	U.S. Geological Survey
USWDS	U.S. Web Design System
VPN	Virtual Private Network
WERC	Western Ecological Research Center
WHCMSC	Woods Hole Coastal and Marine Science Center
WHISPers	Wildlife Health Information Sharing Partnership event-reporting system
WRET	Web Re-Engineering Team
XML	Extensible Markup Language
XSEDE	Extreme Science and Engineering Discovery Environment

U.S. Geological Survey Community for Data Integration 2019 Workshop Proceedings—From Big Data to Smart Data

By Leslie Hsu

Abstract

The U.S. Geological Survey (USGS) Community for Data Integration (CDI) Workshop was held during June 3–7, 2019, at Center Green in Boulder, Colo. The theme of the workshop was “From Big Data to Smart Data” with the purpose of bringing together the community to discuss current topics, shared challenges, and steps forward to advance twenty-first century science at the USGS. The workshop agenda was driven by the needs of the CDI with topics highlighting current resources and technologies that could help attendees in their daily work. Workshop-session categories included enabling integrated science, computing in the cloud, advancing data management, releasing and preserving science outputs, and improving usability and communication. These proceedings provide documentation of the plenary talks, topical-session content and notes, posters, live demonstrations, and attendee comments from the 2019 CDI Workshop.

Introduction

The U.S. Geological Survey (USGS) Community for Data Integration (CDI) is a community of practice that helps its members develop expertise on all aspects of working with scientific data (Hsu and Colasuonno, 2019). The CDI convened a workshop on June 3–7, 2019, at Center Green in Boulder, Colo. The theme of the workshop was “From Big Data to Smart Data,” with the purpose of bringing together the community to discuss current topics, shared challenges, and steps forward to advance twenty-first century science at the USGS. There were 235 in-person attendees and several dozen virtual attendees over 4 days.

The workshop agenda ([Appendix 1, table 1.1](#)) was driven by the needs of CDI members, with topics highlighting resources and technologies that could help attendees in their daily work. Twenty-two topical sessions were led by CDI members and fit into five categories: enabling integrated science, computing in the cloud, advancing data management, releasing and preserving science outputs, and improving usability and communication. Plenary speakers from the community talked about components of integrated science, FAIR (findable, accessible, interoperable, and reusable) data principles, new U.S. governmental acts affecting scientific data, and examples of actionable data in the USGS. The DataBlast poster and live-demonstration session showcased 45 projects from around the CDI and included recent CDI-funded projects as well as initiatives by USGS and partners related to data, software integration, and discovery. Guest speakers from the Science Gateways Community Institute (SGCI, <https://sciencegateways.org/>) were invited to present about value propositions, understanding your audience, and goal setting; and from Earth Science Information Partners (ESIP, <https://www.esipfed.org/>) to present information about connection opportunities within their collaboration areas and funded-project program.

Importantly, the CDI workshop provided a forum for scientists, technologists, data and resource managers, program managers, and others to convene in person to discuss intersecting methods, interests, challenges, and solutions related to scientific data and technologies. “Birds of a Feather” meetings were convened on topics such as software development, diversity and inclusion, DataOps, and drone-observation data. Sharing of ideas from all attendees was encouraged through the use of a mobile application to collect real-time questions and feedback from the audience.

Outcomes of the workshop included key take-aways from the topical sessions and input on the future direction of the CDI. Future areas of collaboration and learning were identified, including data visualization, usability, the importance of software best practices to support scientific research, collaboration with external partners, stakeholder needs, and artificial intelligence (AI) and machine learning methods. The CDI is building on the results of the workshop to guide its future topics, events, and funding opportunities to support an integrated science capacity for the USGS.

2 U.S. Geological Survey Community for Data Integration 2019 Workshop Proceedings—From Big Data to Smart Data

This report is a record of what occurred at the workshop and thus is a reflection of the topics and activities that were important to the CDI at this time. Our intention is that this report will help to guide the future activities of the CDI as well as continue to seed fruitful connections in our diverse community (fig. 1).



Figure 1. Attendees at the 2019 Community for Data Integration Workshop during a break in the plenary schedule. Photograph by Jacob Massey.

Presentations

Abstracts of the plenary presentations and panels are listed here, in order of occurrence. (All speakers are from the USGS unless noted otherwise. Contact information is listed in [Appendix 2, table 2.1.](#))

Welcome and Opening Remarks—From Big Data to Smart Data

By Kevin T. Gallagher and Tim Quinn

The Community for Data Integration has grown over the past 10 years. In 2009, it was a 13-person writing team charged with developing an approach for data-integration goals. Today in 2019, it is a 1,200-person community of practice working together to solve data-integration challenges. The knowledge gained over the past 10 years has positioned the community to contribute to the new USGS vision for integrated Earth-system-characterization science. In this new vision, USGS data products and decision-support tools will be able to consume and model real-time data to generate actionable information.

Turning Your Data Into Real-Time Actionable Insights

By Benjamin Tuttle, Arturo Inc.

This talk highlights the ETL (Extract, Transform, Load) needs for artificial intelligence, machine learning, and deep learning to be successful. It will also focus on the importance of training and labelling to get useful model results and the challenges in this process. Finally, it will discuss the potential for using artificial intelligence and machine learning on real-time spatial data to turn them into real-time insights—that is, turning big data into smart data.

Lightning Talks—Part 1

A Grassland-Productivity Forecasting Tool for the U.S. Southwest, Sasha Reed

Coupling Hydrologic Models in an Interoperable Modeling Framework, Mark Piper, University of Colorado

Integrating eDNA Data Into the USGS Nonindigenous Aquatic Species Database, Jason Ferrante

Tracking Ice Jams With Angular Material and Fargate, Hans Vraga

Emerald Ash Borer, PhenoForecast, 2018, Jake Weltzin

ISO (International Organization for Standards) Metadata-Content-Specifications Workshop, Frances Lightsom

How Can a Standard Be Dynamic? CMECS! Frances Lightsom

Join the New Department of the Interior (DOI) Invasive-Species Data Managers' Community of Practice (COP),
Annie Simpson

U.S. Geological Survey National Digital Trails Network, Greg Matthews

Automated ScienceBase (SB) Data-Release Workflow: A Script to Update Metadata and Populate SB Pages, Emily
Sturdivant

Integrated Modeling: Making Data and Models FAIR With AI and the Semantic Web, Ken Bagstad

Free Chocolate!!! Short Digital Object Identifier Usability Test Required, Madison Langseth

What's New in the ScienceBase Data-Release Process? Tamar Norkin

SageDAT: Data and Tools to Support Collaborative Sagebrush Ecosystem Conservation and Management, Steve Hanser

USGS Cloud Hosting Solutions: Who Are We? What's New? and How Do I Get Started? Jennifer Erxleben

A Next-Generation Graphical User Interface (GUI) for Exploring N-Dimensional Array Data, Rich Signell

Partner Presentations

Data + Community = Action

By Erin Robinson, Earth Science Information Partners

Science funders around the world, both Federal and private, are spending millions of dollars on large-scale, data-intensive, collaborative science projects. After all, the major challenges facing us in all fields from energy to food security to climate-mitigation strategies are not solved by a single person, institution, or domain. These challenges require all of us working together. This talk will explore the relationship between data-intensive science and collaborative community efforts made by organizations such as the Earth Science Information Partners (ESIP) and USGS's CDI to move science forward. Together, if we are better collaborators, we will make data more interoperable and actionable and move science beyond where we thought possible.

Let's Play: How the ESIP Lab Is Supporting Innovation and Ingenuity in Earth-Science Technology Development

By Annie Burgess, Earth Science Information Partners

Not all researchers, developers, and data providers have the resources or expertise to experiment with and deploy new data systems or analysis techniques that push the boundaries of the current technology landscape. ESIP's Lab is working to solve this problem. Through small-grant funding and community input, the ESIP Lab supports projects working towards adoption of community-accepted best practices in scientific-data management, analysis and experimentation with emerging technologies.

Science Gateways Community Institute

By Claire Stirr and Juliana Casavan, Purdue University

Science Gateways are online-community spaces for science and engineering research and education. The Science Gateways Community Institute provides services and resources for gateways, including support for building and running gateways, opportunities for networking and community, and education and training. This presentation provides an overview of services that may be useful to you and an exercise in effectively communicating information about your project by using “napkin drawings” (fig. 2).

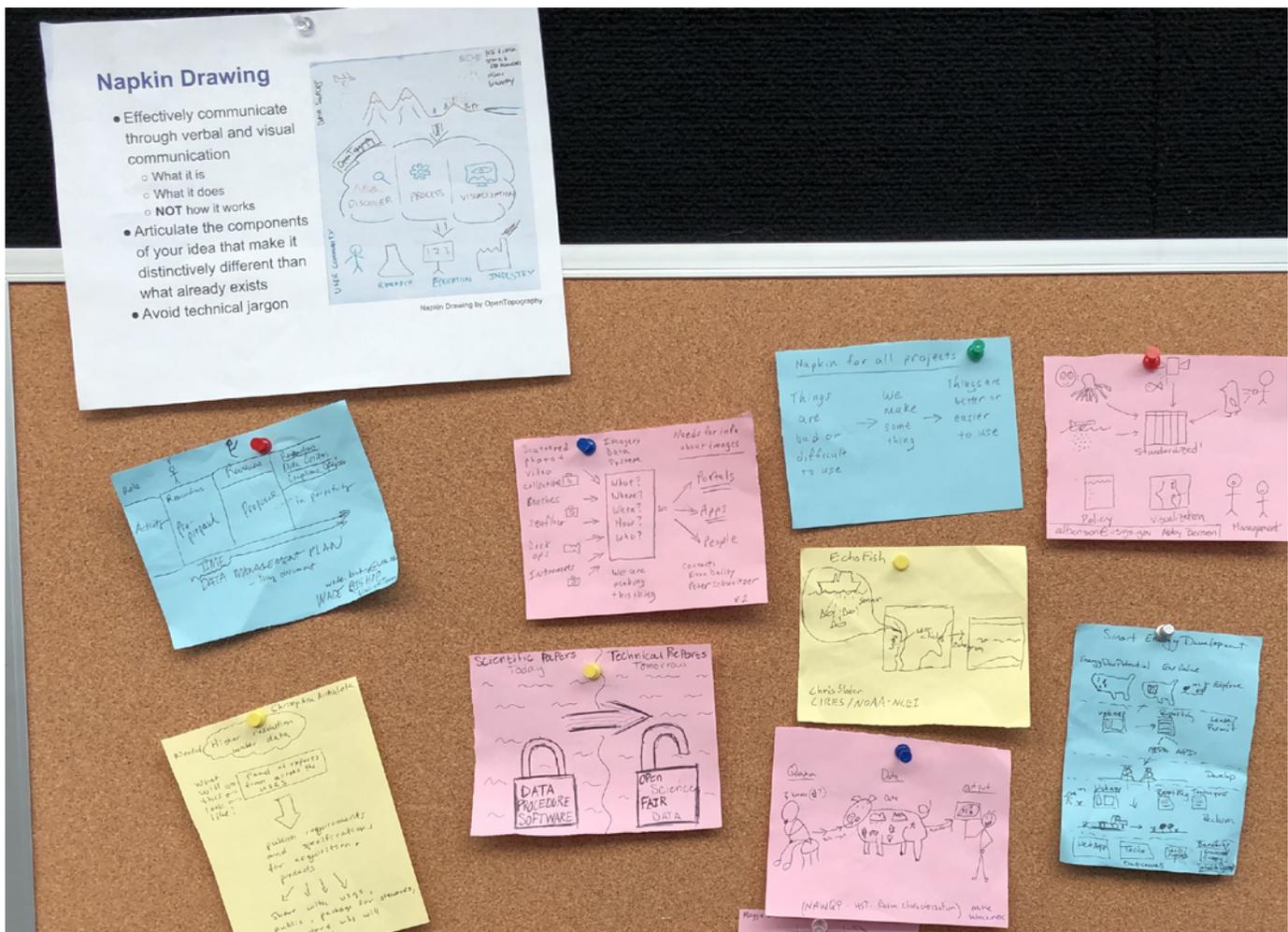


Figure 2. “Napkin drawings” inspired by the presentation by the Science Gateways Community Institute. Photograph by Leslie Hsu.

Components of Integrated Science

USGS National Hydrography Infrastructure

By Sue Buto

In the future, web portals will enable you to search, locate, and discover information related to water by navigating the virtual stream network. To support that future, the USGS National Geospatial Program (NGP) is expanding its focus from individual hydrography and hydrologic-unit datasets to the development of a National Hydrography Infrastructure (NHI). This presentation will provide an overview of the NHI and describe how it will underpin interagency hydrologic observing systems and support integrated science.

USGS Risk Map—A Foundation for Integrated-Risk Science

By Nate Wood

This talk highlights the goals and products of the ongoing collaboration between the USGS and the Department of the Interior (DOI) collaboration to characterize, compare, and map a wide array of hazards to DOI assets across the country. Learn how you can leverage our geodatabase and web services of hazard and DOI-asset data to do additional integrated-risk science.

Advanced Research Computing (ARC): Providing Essential Tools for Integrated Science

By Janice Gordon

This presentation will give a brief overview of the High-Performance Computing (HPC) capabilities and services made available to all USGS and DOI researchers through the Advanced Research Computing (ARC) program. Learn about the different types of HPC systems available for your research and how you can take advantage of training and consulting services.

USGS Cloud-Hosting Solutions Update

By Kimberly Scott

This talk highlights recently released services and features available in the Cloud-Hosting Solutions (CHS) environment as well as upcoming data tools and solutions that will better support integrated science in the future.

“Thinking in Systems”: ScienceBase’s Role in Supporting Integrated Science and Linked-Information Management in the USGS

By Drew Ignizio

Since 2016, ScienceBase has assumed a prominent role in helping to ensure that formally published USGS research outputs are publicly available and hosted in a way to support persistent citation and access. In addition to this function, the ScienceBase team has been carefully structuring how data are stored and establishing programmatic connections to other authoritative systems in the USGS to help ensure that we are able to systematically retrieve and link information to support different data-management needs in the Bureau. Informed in large part from the Configuration Management Committee’s “Linking Systems” summit in 2016, this approach prioritizes standardization, unique identifiers, and the use of Application Programming Interfaces (APIs) to support targeted queries in ScienceBase and to facilitate connections to other authoritative data systems in the USGS.

GeoPlatform

By Tod Dabolt

The GeoPlatform (Geospatial Hosting Platform) supports the discovery, curation, sharing, and use of authoritative geospatial data. It is hosted by the DOI and links decision products backwards through analytics to underlying data. GeoPlatform features include keyword and semantic search, a map viewer, and dynamic digital communities (online spaces to collaborate with colleagues and interact with the public). In addition, GeoPlatform is supporting new approaches to manage and analyze Federal geospatial data in the cloud environment. Find out more at <https://www.geoplatform.gov>.

FAIR Data Session

FAIR 101

By Shelley Stall, American Geophysical Union

The journey toward the FAIR Data Principles (where FAIR stands for findable, accessible, interoperable, and reusable) requires broad support, guidance, and reinforcing policy. The concept of FAIR comes at an inflection point—U.S. Federal policies support openness, new information technology (IT) tools/services hold the potential to revolutionize scientific practice, research funders have introduced mandates and support systems to ensure that the results of the research they sponsor are open to the public, and publishers are adopting open frameworks and strengthening openness requirements for data and methods. In this talk, we will discuss the elements of FAIR and the efforts by funders, publishers, and others to put into place practices and guidance for the cultural changes necessary to make scientific data open and FAIR (fig. 3).

This Is So UnFAIR!

By Jeanne Jones

This talk presents some examples of data challenges that we've encountered during our recent project to identify hazards to people, wildlife, and infrastructure on Department of Interior lands. I can't say that we've seen it all, but we've sure seen a lot of FAIR and unFAIR data. Hopefully, you'll get some tips on how to make your data more user friendly.

Mineral Resources Program: Delivering Data FAIR and Square

By Carma San Juan

The USGS Mineral Resources Program (MRP) collects a wide variety of data to address the Nation's important Earth-science and mineral-resources issues. Part of MRP's mission is to deliver that information to the public. The principles of FAIR emphasize that data should be discoverable, accessible, and usable. These goals are consistent with the goals of MRP, but how do we know how we are really doing when it comes to data delivery and use? This presentation highlights that important question.

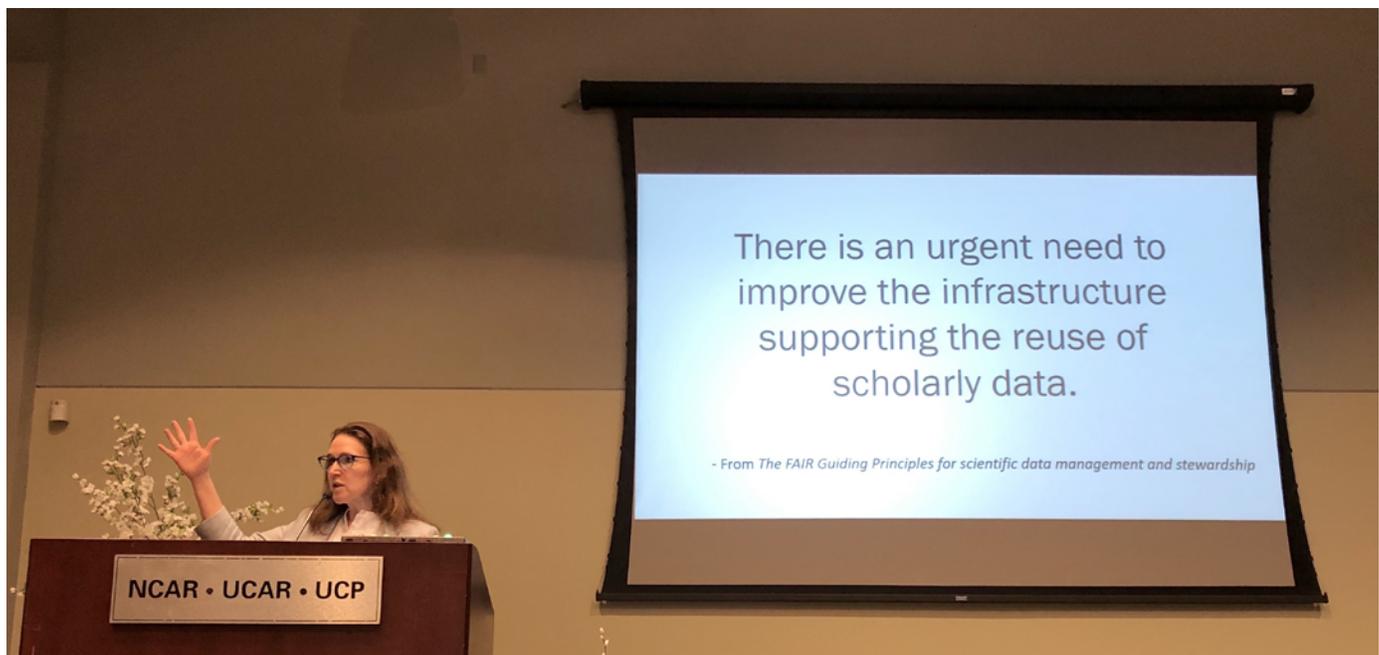


Figure 3. Shelley Stall addressing the plenary audience about FAIR data. Photograph by Leslie Hsu.

Enterprise Tools for Documenting Natural Resources Monitoring Efforts

By Rebecca A. Scully

Collection events for natural-resources data can be easily found and procedures standardized by using [MonitoringResources.org](https://www.monitoringresources.org). These online tools are free to use, are thorough, and encourage collaboration. The goal of the tool is to make data findable and accessible by linking data repositories to projects and protocols and interoperable and reusable with robust documentation of methodologies and designs for long-term monitoring efforts.

Evaluation of Data FAIR-ness From USGS Data Users' Perspectives

By Wade Bishop, University of Tennessee

This presentation assesses data consumers' perspectives on the findable, accessible, interoperable, and reusable aspects of the FAIR Data Principles. Seven USGS inland water scientists and other researchers were asked to think of one of their recent searches for data and describe their processes of discovery, evaluation, and use. Information about data findability and accessibility gives data managers valuable insights into how real-world users access and use data. Results from this study may inform the designs of future data tools, products, and services, as well as indicate which metadata is most needed to make USGS data machine actionable.

Lightning Talks—Part 2

Did Someone Say Speed Data-ing? Bringing the USGS Together Through Science Data Management! Vivian Hutchison

Science for a Risky World—An Introduction to the New Risk Community of Practice, Kris Ludwig

CDI Software Development Cluster—Then and Now, Michelle Guy

Let's Talk Cloud: A Preview—New Frontiers with CHS, Eric Martinez

The New Software Management Website, Cassandra Ladino

Why the International Decade of Ocean Science for Sustainable Development Is Important to the USGS, Sky Bristol

In It for the Long Haul—The National Park Service (NPS) Inventory and Monitoring Division's Long-Term Data

Management Effort, Simon Kingston, National Park Service

Western Ecological Research Center Internet of Things (WERC-IoT): A CHS Pilot Project in Support of North American Waterfowl Research, Cory Overton

The National Land Cover Database (What Is It and What Is It Good For?) Jon Dewitz

Physical Samples: Building Blocks of "Science For a Changing World," Lindsay Powers

Remember, Earth Is a Planet, Too—Bringing Terrestrial Analog Data to the Planetary-Science Community, Marc Hunter

Components of a Data Dictionary, Ray Obuch

Gauging Public Interest in Scientific Information Using Organic Web Search, Tristan Wellman

USGS Big Data to Smart Data

Ecological Forecasting—Making Smart Data From Big Data

By Jake Weltzin

Ecological forecasting, or the prediction of ecological processes and states through space and time, can help resource managers make informed decisions that minimize the adverse impacts of a changing and increasingly variable environment. Recent advances in environmental- data collection and computational power for analysis imply that operational forecasts of ecological processes that were impossible only a few years ago may now be achievable. This presentation describes how iterative, dynamic ecological forecasts can be used to synthesize large amounts of data in decision-support systems that support natural-resource management.

Big Data From a Big Disaster: UAS (Unmanned Aircraft System) Data Collection, Processing, and Dissemination During the 2018 Kilauea Eruption

By Angie Diefenbach

This talk highlights the voluminous amount of multiparametric data collected by UAS during the 2018 Kilauea eruption and the logistics and issues encountered in data management during and following the crisis.

From Big Data to Smart Data in Remote Sensing

By Peter Doucette

The Landsat Program has amassed the longest and most comprehensive record of the Earth's land surface. Petabytes of data have been freely distributed in the last 12 months, resulting in billions of dollars in global economic benefit. Recent investments in producing "analysis-ready data" help users to shift their focus from data preparation to data analysis, which informs decision making. An example is modeling land-use change over time with artificial intelligence and machine learning to better anticipate future land-use changes and inform decision makers.

Process-Guided Deep-Learning Predictions of Lake-Water Temperature

By Jacob Zwart

This talk highlights recent efforts to integrate advanced empirical techniques (deep-learning models) with process knowledge (physics-based models) for the prediction of lake-water temperature. We show that incorporating process knowledge into deep-learning models increased predictive performance when compared to models that used only process-based or empirical techniques. This type of integrated-modeling technique showed the largest performance gains when data were sparse.

Evidence-Based Policy and Department of Interior Data Governance

By Tod Dabolt, Department of the Interior

This presentation describes the Foundations for the Evidence-Based Policy-Making Act, the Information-Quality Act, and the Secretary's Order on Open Science.

Topical Sessions

Brief summaries of the sessions provided here are grouped by category and in order of occurrence. Key take-aways from the sessions are documented in [Appendix 3](#).

Advancing Data Management

Data-Management Workflow Show and Tell

Session moderated by Madison Langseth

Speakers: Dennis Walworth, Ellyn Montgomery, Erika Sanchez, Stephanie Galvan, Meagan Manley, Colin Talbert, Tom Burley, Mikki Johnson, and Danielle Olinger

When it comes to data management, scientists and data managers have a lot on their plates. Science Centers and Programs across the USGS have developed different approaches to managing their data from writing data-management plans and creating metadata with which to publish and preserve their final data products. The goal of this session is to share workflows for data-management tasks and to learn about efficiencies that we can all take back to our respective centers.

Records-Management and Data-Management Connections

Session moderated by Chris Bartlett

What is the relationship between Data Management and Records Management? Do you want to understand how records management affects and connects to data management? The goal of this session is to address the records-management life cycle, its relationship to the data life cycle, and how records are best managed in the USGS. This session is intended to connect records-management requirements to the creation, maintenance, access, and disposition of scientific information, and to explain how the Records Management Program can and does work for and with the science side to address organizing, preserving, and storing scientific records.

We will start by explaining basic terms and concepts, then expand on what is within scope for records management, and how to do it. When do we need to start managing records? How do we do records management? What is a records schedule? Is publishing the same as disposition? How does putting my data in ScienceBase affect what is retained in my organization? If you want the answers to these questions and to learn about available resources, please come to this session.

Advanced Approaches to Data Management: Exploring Services and APIs

Session moderated by Drew Ignizio

Speakers: Marla Hood, Dennis Walworth, Colin Talbert, and Tara Bell

Managing data and business information presents an ongoing challenge for government organizations and private groups alike. The U.S. Geological Survey stores operational information (personnel info, organizational nomenclature, budget information, listings of active research efforts, manuscript and publication listings, catalogs for data products, and so on) across multiple disparate systems. The division of these information holdings is a product of the internal organizational structure of the Bureau, as well as the fact that there is domain expertise among different groups and a division of responsibilities among day-to-day operations. Additionally, some information may be sensitive or have other access restrictions.

Obtaining accurate and complete information, however, often involves pulling content from multiple systems, which can be difficult, or at times, impossible. A long-term solution to these challenges is not another "Silver Bullet" system that attempts to assimilate all of the disparate data holdings, but rather a paradigm shift in how we build these systems to better maintain, access, and integrate the data they store.

Different individuals and groups within the organizational structure often share a need for the same information, albeit to support different workflows for specific tasks. Care must be taken to avoid overfitting a particular process or tool for accessing content from separate data stores; instead, an emphasis must be placed on machine accessibility (APIs) and allowing different groups to pull the appropriate information freely into their own workflows. The solution is "thinking in systems" and putting in place a framework that supports flexible data integration.

The proposed topic is an invitation to a discussion to review some of the core informational systems in the USGS, define best practices for API design and service-oriented architecture, and share notes from those already working on forward-thinking approaches to store, update, and link related high-level information in the Bureau to better support our mission needs.

Content Specifications for ISO Metadata Standard

Speakers: Dennis Walworth, Frances Lightsom, and Joshua Bradley

In this session, we will present findings from the 2018 CDI project "Content specifications to enable USGS transition to ISO metadata standard." The open-ended nature of ISO benefits users with much greater flexibility and vocabulary to describe research products; however, that flexibility means few constraints guide authors and ensure standardized, robust documentation across the USGS. This project provided a proof of concept about how documentation requirements for various USGS products could translate into specifications that, in turn, could be enabled by a metadata-editor user interface and thus guide an author toward robust and standardized metadata by using the ISO international standards. We will demonstrate the initial content-specifications-driven user interface in the mdEditor-metadata authoring tool.

Computing in the Cloud

Cloud Hosting Solutions—Service Offerings in the USGS Cloud

Session moderated by Jennifer Erxleben

Speakers: Eric Larson, Courtney Owens, Rich Signell, Rob Rastovich, and Dionne Zoanni

This session provides a general overview of services currently available or in various phases of development in the CHS environment. Services include Tableau Server, the Cloud Sensor Processing Framework (an Internet of Things (IoT) solution), and High Performance/High Throughput Computing in the cloud. We will describe our "ready-to-use" services, including who is using the service, service configurations, and future directions. The session also highlights several customer-use cases that demonstrate the capabilities of Tableau Server and the Cloud Sensor Processing Framework.

Cloud-Hosting Solutions 101: Demystifying the Cloud

Session moderated by Eric Larson

Speakers: Courtney Owens, Robert Shepherd, and Dionne Zoanni

This session focuses on demystifying common cloud misconceptions by walking through the lifecycle of projects developed in the CHS environment. Join for a general introduction to CHS and to learn more about basic cloud concepts, details for getting started in CHS, the process for proposing new services, and cloud costs and optimization.

Let's Talk Cloud

Session moderated by Michelle Guy

Speakers: Eric Martinez and Jeanne Jones

The cloud offers many resources and services that can be utilized without prior experience with hardware, operating systems, and several third-party tools. Yet it can require a service-oriented, cloud-friendly approach for its users to be successful in utilizing the scalability, capacity, and cost structure of the cloud. We are interested in hearing about how you are using the cloud, what you have learned, what questions you have, what failures and successes you have had in using the cloud. This applies to all scales—from a research application, to real-time processing, to 24/7 operational applications, and everything in between.

Let's Talk Cloud—Applications

Session moderated by Chris Soulard

Speakers: Gabriel Senay, Jessica Walker, and Roy Petrakis

Remote-sensing studies of landscape patterns and changes have historically looked at discrete changes over a small collection of images or have been spatially confined to local-scale analyses. Scientific compromises have been largely driven by storage and computation limitations. Google Earth Engine and other cloud computing environments have ushered in a new era during which the land surface can be evaluated continuously over large spatial areas and long periods of time.

How are USGS scientists using the cloud to conduct research? This session will use a series of case studies to explain how USGS researchers are moving past desktop-driven computing and adopting cloud-based solutions for scientific analyses. By sharing use cases and discussing preferred workflows, the session will aim to facilitate greater use of emerging cloud-computing technologies (like Google Earth Engine) and enhance access to geospatial analyses without requiring the use of traditional geographic information system (GIS)/remote sensing software.

Enabling Integrated Science

Packaging Scientific Analyses as Software

Session moderated by Steve Aulenbach, Abby Benson, Sky Bristol, and Mark Wiltermuth

Speakers: Rich Signell, Jessica Walker, Annie Simpson, Caitlin Andrews, Daniel Wiefelich, and Abhijeeth Baregal

Scientific analyses that are openly available and well documented facilitate communication regarding research progress and collaborative effort, both of which benefit scientific research. Systematizing the process of packaging the results of research into a cohesive construct that can be explained, reviewed, and deployed in a variety of venues can turn scientific analyses into operational decision-analysis capabilities. The packaging includes identifying the provenance of data, documenting the appropriate context and limitations of an analysis, describing the software code used to conduct the analysis, and describing the graphical or tabular output that depicts the key scientific finding of the analysis. Feedback from decision making stakeholders is an aspect of continuously improving analysis packages until they can become useful operational tools. We plan to provide examples of packaging scientific analyses as software and to lead a discussion on these ideas.

Integrating Data and Models for Next-Generation Predictive Science

Session moderated by Ken Bagstad and Mark Piper

Integrated collaborative modeling is increasingly recognized as a key long-term goal for the USGS and other science practitioners. It is also a key purpose and outcome of the FAIR data principles; however, abundant challenges exist to making data and scientific models available and interoperable, including (1) semantics, (2) data storage and services, (3) the use of multiple programming languages and modeling paradigms (both deductive and inductive approaches, including machine learning), (4) uncertainty propagation, and (5) ensuring data and model reuse at appropriate spatiotemporal scales. Modeling protocols and frameworks, like those of the Community Surface Dynamics Modeling System (CSDMS) and Artificial Intelligence for Ecosystem Services (ARIES)/Integrated Modelling Partnership, provide examples of practical methods to overcome these integration challenges. In this session, presenters will address theories and methods to overcome the aforementioned challenges, potential applications, along with potential lessons learned that could inform integrated modeling for the USGS and beyond.

Exploring the Landscape of Scientific-Computing Tools for Large-Scale Analytics, Machine Learning, and Integrated Modeling

Session moderated by Jeff Falgout and Janice Gordon

This session will take a deeper dive into the architecture and use cases for each type of computing resource. The session will cover the currently available Yeti HPC system, the BlackPearl converged-storage system, and the Globus large data-transfer pilot project. We will also discuss the new Cray AI system “Tallgrass” and the new HPC system “Denali,” both coming online in 2019.

Visualization Tools You Can Use: CDI Risk Map, TerriaMap, Tableau

Speakers: Jeanne Jones, Rich Signell, Shayne Urbanowski, Kevin Henry, Dionne Zoanni, and Jason Sherba

Tools for visualizing, comparing, and analyzing our data are numerous and diverse. How do you know what tool is best for your purpose? The CDI Community for Data Integration has supported projects and presentations that evaluate visualization tools and their features. In this session, there will be overviews and demonstrations of visualization tools that have been used in recent CDI and USGS projects, and a discussion of how we would like to collectively learn about visualization capabilities in the future.

Improving Usability and Communication

Improving Collaboration Experiences Between Data Providers and Curators

Session moderated by Sophie Hou

Data providers and curators are two key stakeholders of data-management activities and services from repositories. Successful collaborations between the two roles is not only vital for the productivity of the repositories, but also for achieving data-management requirements and realizing open data. During this session, we are interested in sharing and discussing challenges, solutions, and lessons learned in the following areas: (1) key scenarios that motivate the data providers and curators to collaborate, (2) finding effective balance points between independent and collaborative activities, and (3) data-management tools and data services that facilitate the collaborations.

Want to Know How to Make Your Data Applications Delightful? Usability Can Help!

Session moderated by Sophie Hou and Madison Langseth

Speakers: Tamar Norkin and Lindsay Platt

Are you providing applications that enable users to find, explore, or work with data? Do you know what your users would say about using your applications? If you answer “yes” to the first question but “no” to the second, you will want to join this session!

During this session, the attendees will have the opportunity to (1) learn about the usability techniques that can be applied to desktop and web applications, (2) see usability- evaluation examples and results, and (3) practice how to assess the usefulness of data applications.

SGCI Communication Session—Building your Value Proposition

Session moderated by Juliana Casavan and Claire Stirm

Why would someone be interested in your project? Build a concise value statement that will inform your audience and retain their interest (instead of hitting the back button).

Releasing and Preserving Science Outputs

Preservation and Digitization of Physical Materials

Session moderated by Frances Lightsom

USGS science makes use of physical samples, including rocks, fossils, and biological specimens. USGS scientific data has been recorded on physical objects, including paper, photographic film, sound recordings, and floppy disks. As data managers, we try to preserve physical materials that have continuing value. At this session, a panel of speakers will address the benefits, challenges, and requirements for preserving and digitizing physical materials; share some good examples; and discuss questions with session participants.

Progress on Handling Large Data in the USGS and What's Ahead

Session moderated by Drew Ignizio

Speakers: Jeff Falgout, Matt Davis, Doug Schuster, and Rich Signell

Research organizations may work in different topical areas and use varied tools and approaches in their day-to-day operations, but one thing is true for everyone in the science and research domain: data files are getting bigger. Whether it is the size of dataset inputs or the resulting output that a particular analysis produces, authors and data managers are often hastily playing catch-up in response to today's data storage and access demands. How is this affecting the USGS? What are the current trends and the latest developments in data storage available to our researchers? Does the "Click to Download" model still work for the data we are producing, given the sizes of our products and the workflows of other researchers? This session will provide an opportunity for an update on the current capabilities provided by the Core Science Systems mission area to support USGS scientists, as well as an open discussion for anyone dealing with large data challenges.

Data and Metadata Review and Creation

Session moderated by Frances Lightsom

The USGS Metadata Reviewers Community of Practice is hosting an opportunity to learn about best practices for review of data and metadata, as well as tips, tricks, and guidelines for creating good metadata records and data releases. We hope you will join us and bring questions. We plan to reuse questions and answers to create responsive online guidance, so bring questions other people are asking you as well as your own questions.

Software-Release Questions Answered

Session moderated by Laura DeCicco, Cassandra Ladino, and Eric Martinez

Preparing and managing software is increasingly recognized as a crucial step in the data-management and scientific process. Recently the official USGS software-management and release process has been updated, but concrete answers and resources may yet seem difficult to find. In this session, a panel of USGS experts will field questions related to the policy and provide practical guidance tailored to your specific situation. Come with questions, use cases, or scenarios, and your questions will be answered.

Collaboration Area Meetings

Citizen-Centered Innovation Working Group

Session moderated by Sophia Liu

The Citizen-Centered Innovation Collaboration Area welcomes the participation of anyone interested in Open Innovation efforts like Crowdsourcing, Citizen Science, Civic Hacking, and Challenge and Prize Competitions. The purpose of this community is to

- promote an understanding of the role and potential benefits of open-innovation efforts for the USGS as well as other DOI bureaus and offices;
- facilitate and enhance connections between the USGS and the larger Federal and public Citizen Science and Open Innovation communities;
- provide access to information and tools to support the proper, effective, and creative use of these open-innovation efforts within the USGS and the other DOI bureaus; and
- encourage citizens to foster and improve their scientific literacy and use of USGS and DOI products and services.

During the CDI Workshop, this session will focus on discussing the upcoming efforts to develop a bureau-wide strategy for open innovation with an emphasis on crowdsourcing and citizen science as well as a DOI Generic Information Collection Request to reduce the policy barriers for leveraging these participatory techniques.



Figure 4. Speed data-ing at the Data Management Working Group breakout session. Pairs of attendees discuss their common interests and challenges in data management. Photograph by Vivian Hutchison.

Data-Management Working Group

Session led by Vivian Hutchison and Madison Langseth

What are some good things happening in science-data management that I could really use in my Center or Program? What do we do well in my Center or Program that I could share with others? Join the CDI Data Management Working Group to discover data-management tips, tricks, workflows, and other activities of interest across the Bureau as we build our agenda of topics for the coming year. Come and meet your colleagues, who are all working towards common goals to intelligently manage science data ([fig. 4](#)).

Software Development Cluster

Session led by Blake Draper, Michelle Guy, and Cassandra Ladino

Speakers: Tom Burley, Laura DeCicco, Mike Hearne, Dell Long, and Aaron Stephenson

This session features a panel of USGS employees who will address the topic of diversity in software and what “software” means to them. Smaller groups then will discuss opportunities for cross-USGS collaboration on software development, the USGS Software Informational Memo, USGS institutional support for software development, and USGS opportunities for the career growth of software developers.

Birds of a Feather Discussions

Birds of a Feather discussions are informal meetings about a particular topic of interest to the CDI community (fig. 5). The sessions were proposed and held during the workshop.

Title	Contact	Time Date	Place
Software Development Collaboration Area	Michelle Guy	Thursday Lunch 12:30pm	Here
Tech Stack Collaboration Area			
Diversity + Inclusion in the USGS	Marcia McNiff (Diversity Change Agent)	Thurs (6/16) Lunch	Residence Inn Marriott Lobby
Water MA meet-and-greet	Linda DeBruin	6/5 lunch BYOL	Res. Danc Inn Marriott Lobby
DataGPS	Rich Signell	WED 12:30 LOBBY	
UAS / DRONES / DATA	JOE ADAMS	THUR. 12:30	NORTH AUDITORIUM

Figure 5. List of Birds of a Feather discussions during the week. Photograph by Leslie Hsu.

Trainings

CDI Workshops provide trainings that help attendees improve their skills in working with data.

Introduction to R

Contact: Lindsay Platt

This course is intended for novice R users (including those with absolutely no experience). This course is focused on basic programming skills and best practices, as well as tools for scientific workflows (publication-ready plots, statistics, and USGS packages). Methods for accessing, cleaning, analyzing, and visualizing USGS data are reviewed and practiced. Our main goal is for participants to leave with the confidence to incorporate R into their own workflows. The curriculum is adapted from the full 3-day course, which is available at <https://owi.usgs.gov/R/training-curriculum/intro-curriculum>. The course is interactive and hands-on, and requires students to attend with a laptop computer with R software installed (instructions available at <https://owi.usgs.gov/R/training-curriculum/installr>).

USGS Software-Release Practicum

Contact: Lance Everette

Ninety-nine software releases have been disseminated since the implementation of the USGS software release policy (Instructional Memorandum OSQI [Office of Science Quality and Integrity] 2016-01; note that this policy was replaced by Instructional Memorandum OSQI 2019-01 [<https://www.usgs.gov/about/organization/science-support/survey-manual/im-osqi-2019-01-review-and-approval-scientific>] after the workshop on October 2, 2019). By comparison, the USGS disseminated 6,488 publications during the same period. Given USGS policy and how critical software is to modern research, these two data points are hard to reconcile. An often discussed obstacle to releasing USGS software is the perceived difficulty of reviewing and approving large, complex software products; however, passive monitoring of USGS products by the USGS Web Re-Engineering Team (WRET) indicates that large software products are not necessarily the USGS norm, and the experience of the CDI Data at Risk project illustrates that small, simple code sets can be relatively simple to review, approve, and release.

This “Software Release Practicum” intends to instill confidence in the software-release workflow by walking participants through a simple USGS software release in real time, resulting in a new disseminated USGS Software Release. This practicum will provide a software-release training exercise and a revealing discussion of the intersection of USGS scientific culture and software policy.

We propose that the practicum work through the following three stages of a software release: Software Project Planning and Development, Software Reviews, and Dissemination/Release.

ISO Metadata-Content Specification Workshop

Contact: Dennis Walworth and Frances Lightsom

In this workshop we intend to gather people from across USGS mission areas as well as interested collaborators to build on the work initiated by the 2018 CDI project “Content specifications to enable USGS transition to ISO metadata standard.” The open-ended nature of ISO benefits users with much greater flexibility and vocabulary to describe research products; however, that flexibility means few constraints to guide authors and ensure standardized, robust documentation across the Bureau. This project provided a proof of concept of how documentation requirements for various USGS products can translate into specifications that in turn can be enabled by a metadata-editor user interface, thus guiding an author toward robust and standardized metadata that meets the ISO international standard.

During this workshop, we will continue work on common documentation modules already developed—Basic, Geospatial, Lineage, Taxonomy—and build on these with Observational, Experimental, and Computational modules that describe more specifically the various types of data produced by the Bureau. Gathering and reviewing metadata user stories will be an important part of our discussion. This workshop will provide us with an opportunity to define a metadata approach that is based upon what we need to document—both common and domain-specific aspects of our data—and to move away from a one-size-fits-all “standard” driving our data documentation. Knowledge of ISO metadata is not needed; however, some knowledge of the Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata is helpful.

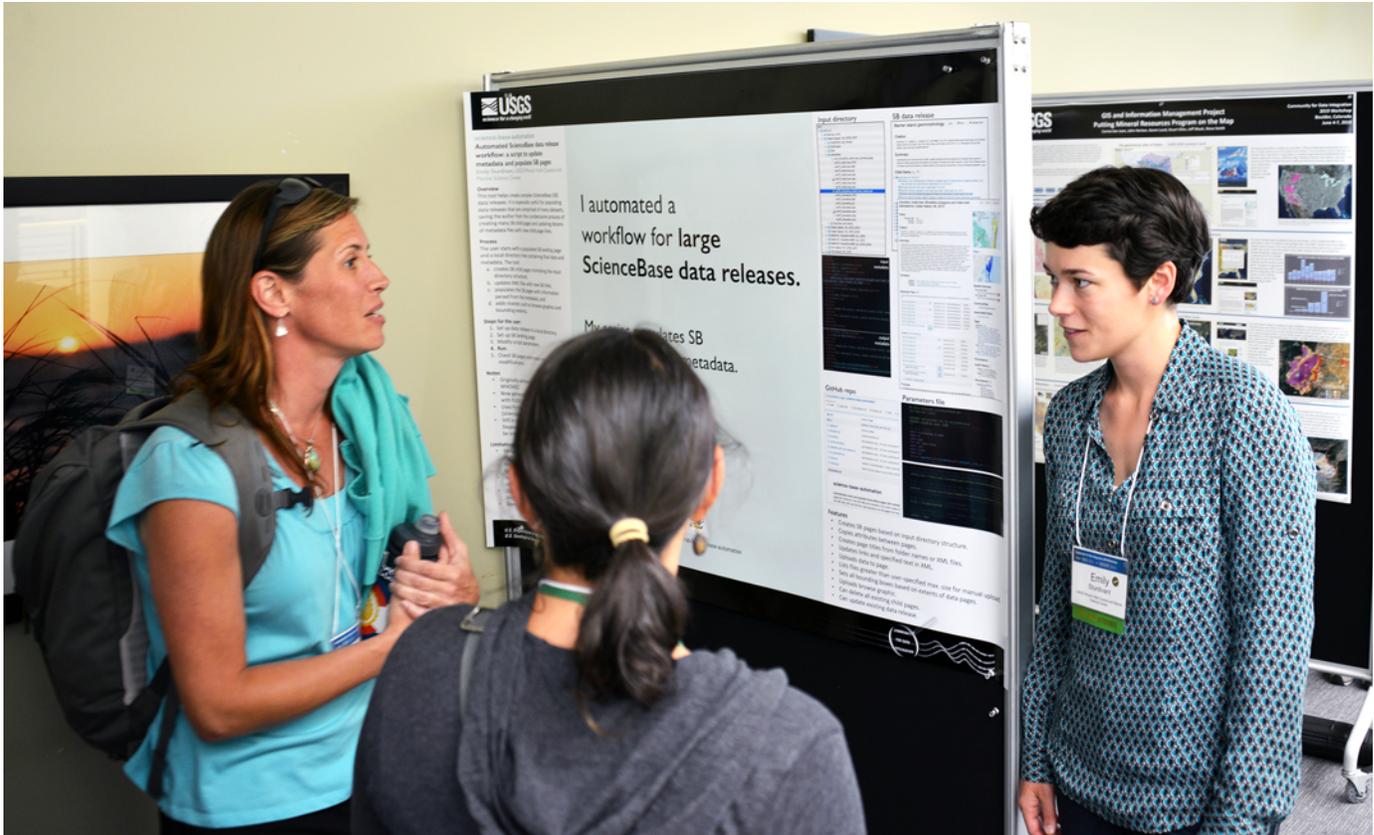


Figure 6. Meeting attendees discuss their work at the DataBlast poster and demonstration session. Photograph by Jacob Massey.

High-Performance Computing

Contact: Janice Gordon

This half-day course provides an introduction to USGS high-performance computing systems and describes how to use them effectively. We will first provide some background on high-performance computing systems, then demonstrate how to login to the USGS Yeti system from your computer, and finally describe how to submit scripts and run jobs. Hands-on exercises require a USGS laptop with Virtual Private Network (VPN) capabilities.

DataBlast

The DataBlast is an informal poster and live-demonstration session designed to ignite creative discussions and build community (fig. 6). All workshop attendees were invited to participate and share information about the projects on which they were working. The abstracts below were submitted by attendees prior to the DataBlast. Each abstract identifies in brackets the CDI Science Support Framework category (Appendix 5) with which it is most aligned. The following abstracts are arranged in alphabetical order by the primary author's last name.

USDA (U.S. Department of Agriculture) Forest Service National Riparian Inventory Base Map

By Sinan Abood (U.S. Department of Agriculture Forest Service), Michael Wiczorek (USGS), and Linda Spencer (U.S. Department of Agriculture Forest Service)

[Science Support Framework category: Applications]

Riparian ecotones are an important natural resource with high biological diversity. These ecosystems contain specific vegetation and soil characteristics that support irreplaceable values and multiple ecosystem functions and are very responsive to changes in land-management activities. Delineating and quantifying riparian areas is an essential step in riparian monitoring,

planning, management, and policy decisions. The USDA Forest Service supports the development and implementation of a national-context framework with a multiscale approach to define riparian areas by using freely available national geospatial data. Riparian Buffer Delineation Model (RBDM) v5.2 in Esri ArcGIS Pro is used to process all the input data. RBDM recognizes the dynamic and transitional nature of riparian areas by accounting for hydrologic, geomorphic, and vegetation data as inputs. Here we present a new a national variable-width base map of riparian areas at a 10-meter spatial resolution for all NHDPlus 1:100K streams and the associated modeling technique that was used to create the map.

New Imagery Data System for the Coastal/Marine Hazards and Resources Program

By Seth Ackerman, Evan Dailey, Heather Schreppel, Peter Schweitzer, and Frances Lightsom (all USGS)
[Science Support Framework category: Publishing/Sharing]

Imagery provides valuable information about the processes, structures, and composition of the coasts and sea floor. The Coastal and Marine Hazards and Resources Program (CMHRP) houses expansive holdings of oblique aerial, seafloor, and drone imagery and videos. They are stored inconsistently, however—some in the CMHRP’s Video and Photograph Portal, but most in published reports and data releases in difficult-to-use formats (often as zipped archives). Video data are often distributed in even less user-friendly formats.

To solve these challenges, CMHRP is creating a new centralized imagery database—the Imagery Data System. It is planned to include: (1) a database housed in the CMHRP-trusted digital repository; (2) standards and procedures for data and metadata ingestion; (3) access restrictions for imagery flagged as provisional; (4) an API to provide imagery to CMHRP projects, including portals that present CMHRP science to external customers (for example, CMHRP’s Coastal-Change Hazards and Video and Photograph Portals); and (5) an internal interface to locate and download data files.

Preservation and Publication of Historical Seismic Data from the U.S. Geological Survey

By Matt Arsenault (USGS)
[Science Support Framework category: Science Data Lifecycle—Preservation]

The USGS Woods Hole Coastal and Marine Science Center (WHCMSC) has actively collected geophysical data in the Pacific Ocean, Gulf of Mexico, and Atlantic Ocean for several decades, including both shallow subsurface-reflection data and deeper “blue water” reflection data. Prior to the early 1990s, most geophysical data were collected in analog format as paper rolls showing continuous profiles up to 25 meters (m) long. WHCMSC currently holds hundreds of geophysical surveys that are vulnerable to degradation or loss in its paper repository and are inaccessible to most researchers because they are not available digitally. As part of a bureau-wide effort, WHCMSC is scanning its analog geophysical holdings into TIFF (Tagged Image File Format) images. From the TIFF format, they are planned to be converted to industry standard SEG–Y format by using open-source software and distributed to the public by using The USGS Coastal and Marine Geoscience Data System (CMGDS). Access to these resources will help to inform future surveys, especially by preventing costly redundant marine geophysical surveys. Early efforts to rescue these historical data have led to planning more efficient data collection in the deep Atlantic, reducing the number of survey lines needed in a study of the Chesapeake Bay, and providing subsurface geologic information to correlate with modern surficial data from Lake George, N.Y.

Transforming Biosurveillance by Standardizing and Serving 40 Years of Wildlife-Disease Data

By Neil Baertlein, David Blehert, LeAnn White, and Ali Rahama (all from USGS)
[Science Support Framework category: Data Management]

During the past 40 years, the USGS National Wildlife Health Center (NWHC) has amassed the world’s largest repository of wildlife-disease-surveillance data. Prompted by the development and acquisition of new data systems (WHISPers [Wildlife Health Information Sharing Partnership event-reporting system] and LIMS [Laboratory Information Management System]), as well as the desire to apply FAIR data principles, the NWHC has proposed a framework to identify, document, and prepare its diagnostic datasets for being archived or migrated to new systems. This framework has been laid out in a five-step process: 1. Definition, 2. Classification, 3. Prioritization, 4. Cleansing and Standardization, and 5. Mapping and Migration. Completion of this process is planned to enable NWHC to supply summary information (for example, cause of morbidity/mortality, location, species) to the Wildlife Health Information Sharing Partnership event-reporting system (WHISPers), which is a publicly accessible online repository for current and historical information on morbidity and mortality events affecting wildlife in North America. These data are planned to allow our partners (for example state agencies and the Federal DOI and Department of Homeland Security agencies) to access and leverage these data for assessing disease threats, providing situational awareness, and developing predictive capabilities.

Integrated Modeling: Making Data and Models FAIR With AI and the Semantic Web

By Ken Bagstad (USGS), Ferdinando Villa (Basque Centre for Climate Change), and Stefano Balbi (Basque Centre for Climate Change)

[Science Support Framework category: Applications]

Interoperability among varied multidisciplinary data and models based on the FAIR data principles, combined with AI, offers a path forward to make scientific modeling faster and more transparent, while improving the reuse of existing knowledge. This poster describes how the ARIES (Artificial Intelligence for Ecosystem Services) environmental and Earth-sciences-modeling platform fosters interoperability among a wide range of scientific data and models while applying AI (machine learning, machine reasoning, and semantics) to fully support both deductive and inductive modeling. Since 2007, the ARIES team has worked to develop (1) semantics that work across diverse scientific disciplines and describe data and model elements with the underlying logic and parsimony to support machine reasoning; and (2) open data and models that can be linked together by using (3) open-source software tools. These software tools include a technical ARIES Modeler interface that experienced scientists and modelers can use to link and contribute data and models to a growing knowledge base. Additionally, a web-based ARIES Explorer allows nontechnical stakeholders to run models, visualize data, download model input and output data, and understand model-workflow provenance. A concurrent demonstration of the ARIES Modeler and Web Explorer will be provided at the workshop.

Integrating Short-Term Climate Forecasts Into a Restoration-Management Support Tool

By John Bradford, Caitlin Andrews, David Pilliod, Justin Welty, Michelle Jeffries, and Linda Schueck (all USGS)

[Science Support Framework category: Applications]

Land managers are regularly required to make decisions regarding restoration, often based on little more than business-as-usual practices and in the face of future climate uncertainty. In the dryland, temperate ecosystems of the United States, restoration success is largely dictated by the extent to which severely dry surface-soil conditions can be avoided so that plants can develop sufficient root systems. Although recognition of these risks is growing, no tools are available to help natural-resource managers understand short-term, site-specific exposure. A new product from the National Weather Service provides short-term (less than 1 year) forecasts of future climate conditions. We will leverage the established Land Treatment Exploration Tool (<https://chsapps.usgs.gov/apps/land-treatment-exploration-tool/>) and the framework of our 2018 CDI-funded Long-Term Drought Simulator to create a new product that will enable managers to quickly access multimonth climate forecasts and provide quantitative guidance about what these forecasts mean for the probability of seeding and planting success. In addition, we will showcase our Long-Term Drought Simulator as an example of the power of exploring site-specific drought trends and predictions.

Developing Modern Web Pages by Using the U.S. Web Design System

By Mary Bucknell and Jim Kreft (both from USGS)

[Science Support Framework category: Information]

[Waterdata.usgs.gov](https://waterdata.usgs.gov) is the third most-visited website in the Department of Interior. Over the past year, we have begun work on the next generation of this critical data-dissemination system, starting with the most popular part: pages that give real-time water data for thousands of monitoring locations across the country. Core requirements are that the pages first be mobile and then accessible, interactive, and modern with respect to their look and feel. The use of the U.S. Web Design System (USWDS, <https://designsystem.digital.gov/>) was fundamental to this effort. The system provides both written guidance and style sheets that can be used to produce a website that meets USGS requirements, while still allowing for customization of design elements such as color. The beta version of the site pages has been released to the public, and the code behind the site pages is available to the public at <https://github.com/usgs/waterdataui>. We will describe how we used the USWDS to create our own Water Data for the Nation visual-ID package to meet USGS visual ID requirements, how we are approaching accessibility for time-series graphs, and how we are using tools to determine the level of accessibility of the pages.

A Community Supporting a Community: How CDI Members Can Use ESIP Lab Technology Development, Sharing, and Evaluation Services

By Annie Burgess (Earth Science Information Partners)

[Science Support Framework category: Science Project Support]

The ESIP Lab supports Earth scientists and Earth data technologists build insightful tools to search, serve, analyze, and visualize data that describe the Earth. Through small grants, community input, and strategic outreach, the ESIP Lab provides a unique suite of services missing in the science-innovation sector. This poster will show how the ESIP Lab's combination of community building and technology development can help the CDI and the USGS fulfill their missions.

Ice Jams—Mobile-Friendly Website for Collection and Display of Real-Time and Historical Ice-Jam Information

By Kathy Chase, Lauren Privette, Jeremy Newson, Hans Vraga (all USGS)

[Science Support Framework category: Applications]

Ice jams along rivers cause flooding, scouring, structural and environmental damage, and injuries and loss of life; thus, 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 later threaten lives and property. This project will enable individuals to collect real-time information about ice jams.

ScienceBase and EROS Collaboration to Release UAS Imagery

By VeeAnn A. Cross, Sandra M. Brosnahan, Drew Ignizio, and Ryan Longhenry (all USGS)

[Science Support Framework category: Science Data Lifecycle—Publishing/Sharing]

To efficiently and effectively meet the increased demands for USGS datasets, manage large data volumes, minimize data-storage redundancy, and adhere to the USGS Survey Manual data-publishing requirements (Survey Manual 502.8, <https://www.usgs.gov/about/organization/science-support/survey-manual/5028-fundamental-science-practices-review-and>), USGS ScienceBase and the Earth Resources Observation and Science (EROS) Center are collaborating to provide access to and distribution of USGS UAS imagery. ScienceBase and EROS are trusted digital repositories (a USGS requirement for the preservation of digital scientific data; <https://www.usgs.gov/about/organization/science-support/office-science-quality-and-integrity/acceptable-digital>), yet each plays a unique role in data display, discovery, access, and distribution. ScienceBase (<https://www.sciencebase.gov>) is a collaborative scientific database that provides data cataloging and management and is a primary platform in which USGS scientists can publish data. ScienceBase also meets USGS data-release requirements, which include providing compliant metadata and perpetual landing pages for DOI universal resource locators (URLs). EROS EarthExplorer (<https://earthexplorer.usgs.gov>) provides a user interface that enables online search, discovery, and download of remotely sensed imagery as either individual images or a complete dataset. The ScienceBase and EROS collaboration will use EarthExplorer as the distribution mechanism for UAS imagery published through ScienceBase. This collaboration leverages the strength of each entity, increases the discoverability and the size of the user community for USGS scientific data, and follows good data-management practice by reducing data-storage redundancy.

Have You SEINeD Your Data Lately? A National Public Screening Tool for Invasive and Nonnative Aquatic-Species Data

By Wesley Daniel (USGS), Matthew Neilson (USGS), Ian Pfingsten (Cherokee Nations Technologies, contractor to USGS), Craig Conzelmann (USGS), Cayla Morningstar (Cherokee Nations Technologies, contractor to USGS), and Justin Procopio (Cherokee Nations Technologies, contractor to USGS)

[Science Support Framework category: Web Services]

Identifying the leading edge of an invasion can be difficult, especially when the invader is a native transplant or is not well known. Many management and research groups have conducted biological surveys that may contain unintentionally collected, unrecognized data on nonindigenous species. These groups need a tool to help them screen large amounts of data for the identification of nonnative species. The USGS Nonindigenous Aquatic Species (NAS) Database (<https://nas.er.usgs.gov/>) team proposes a novel publicly available tool that has been developed to screen invasive and nonnative species data. The SEINeD (Screen and Evaluate Invasive and Nonnative Data) tool is planned to allow the users to upload a biological dataset that can be

screened for any of the more than 1,290 invasive or nonnative species of fish, invertebrates, amphibians, reptiles, mammals, or plants that the NAS program tracks. The user will then be informed if the species is within its native range or has been introduced. This effort will represent the newest national tool for early detection and rapid response for state and Federal management efforts.

NHDPlus High Resolution

By Ariel Doumbouya, Karen Adkins, Ellen Finelli, Becci Anderson, Alan Rea, and Hayley Thompson (all USGS)
[Science Support Framework category: Data]

The National Hydrography Dataset Plus High Resolution (NHDPlusHR, <https://www.usgs.gov/core-science-systems/ngp/national-hydrography/nhdplus-high-resolution>) is an integrated geospatial-data product that incorporates the National Hydrography Dataset (<https://www.usgs.gov/core-science-systems/ngp/national-hydrography/national-hydrography-dataset>), data from the 3D-Elevation Program (<https://www.usgs.gov/core-science-systems/ngp/3dep>), and the Watershed Boundary Dataset (<https://www.usgs.gov/core-science-systems/ngp/national-hydrography/watershed-boundary-dataset>). NHDPlusHR is currently (2019) being produced and distributed by the U.S. Geological Survey National Geospatial Technical Operations Center. NHDPlusHR data provide all of the NHDPlus Version 2 attributes such as natural-flow estimates, flow adjustments for diversions, stream order, and much more, with the additional detail of the current High Resolution NHD and 1/3-arc-second seamless digital-elevation models.

High-Resolution Interagency Biosurveillance of Threatened Surface Waters in the United States

By Sara L. Eldridge, Elliott Barnhart, and Adam Sepulveda (all USGS)
[Science Support Framework category: Science Data Lifecycle—Processing]

Standardized environmental DNA (eDNA) detection techniques have the potential to provide biological observations matching the scale and quality of in situ physical and chemical measurements. Successfully joining these disparate data streams would support comprehensive assessments of ecosystem stressors such as pathogens, invasive species, or harmful algal blooms. In cooperation with the Monterey Bay Aquarium Research Institute, USGS scientists have incorporated a portable robotic environmental-sample processor at USGS streamgages to collect near-real-time DNA (deoxyribonucleic acid) biosurveillance data. Here, we are planning to expand this work by developing a data-science pipeline for processing covariate datasets through sequential chained steps for conducting ecological-risk assessments. Using eDNA samples and data collected from the Yellowstone River in 2018, we plan to develop spatiotemporal, multiscale occupancy models to identify how changes in streamflow and weather/climate data are associated with the presence or absence of *Tetracapsuloides bryosalmonae*, *Salmo trutta* (brown trout), and *Escherichia coli*. The planned computationally expedient R package may be used to fit time-series-informed, multiscale occupancy models with environmental covariates by using Bayesian computation methods. Planned key outcomes are an interagency data-sharing framework for improving ecological forecasting, assessments of biological hazards, and a data-science pipeline that can be customized to integrate similar data streams for different biosurveillance applications.

Develop Cloud Computing Capability at Streamgages Using Amazon Web Services Greengrass IoT Framework for Camera Image Velocity Gaging

By Frank Engel (USGS)
[Science Support Framework category: Web Services]

The USGS Surface Velocity Workgroup (SurfBoard) is developing and testing computerized video-based approaches to measuring streamflow during floods from video-derived stream velocities (image velocimetry). Often, hydrographers cannot safely measure streamflow data at streamgages during floods because of dangerous site conditions or event timing. Image velocimetry offers a solution to this safety issue because it allows hydrographers to measure streamflow remotely. To support progress toward achieving this goal, the SurfBoard has partnered with USGS Cloud-Hosting Solutions to develop an Internet of Things (IoT) provisioned-image streamgage that applies existing equipment and processing elements to edge computing by using the Amazon Web Services (AWS) IoT and Greengrass Core software. The work has two objectives: (1) prepare and test an IoT framework that replicates the existing image of streamgage flows in the AWS Cloud; and (2) translate selected processing algorithms into cloud-based programs (AWS Lambda functions) within the IoT framework.

Insights gained are being used to build a decision matrix aimed at leveraging IoT applications for other streamgages and sensors in the network. Image-velocimetry measurements based on an IoT approach are being tested at two streamgages.

USGS Cloud-Hosting Solutions Support 21st-Century Science

By Jennifer Erxleben (USGS)

[Science Support Framework category: Science Project Support]

CHS provides a Cloud-based enterprise environment that supports hosted applications and a collection of managed services to advance the mission of the U.S. Geological Survey (USGS). By leveraging cloud technologies, CHS is enabling the USGS to operate as a 21st-century organization. CHS is developing an ecosystem of cloud-based services and tools that allow scientists to acquire, secure, store, share, process, analyze, model, visualize, and disseminate data in a common environment. CHS is operated and maintained by USGS Enterprise Information staff in cooperation and governance with the Department of the Interior.

Establishing Standards and Integrating Environmental DNA (eDNA) Data Into the USGS Nonindigenous Aquatic-Species Database

By Jason Ferrante, Matthew Neilson, Wesley Daniel, and Margaret Hunter (all USGS)

[Science Support Framework category: Data Management]

Nonnative species continue to spread in sensitive ecosystems such as the Florida Everglades and the Great Lakes Basin; identifying the presence of these species in such large, remote systems can be challenging. Environmental DNA (eDNA) provides an opportunity to perform high sensitivity monitoring efforts of cryptic species by testing water, soil, or air samples for DNA. We aim to develop a mechanism for adding eDNA data to the NAS database (<https://nas.er.usgs.gov/>). The database currently maps and displays the distribution of nonnative aquatic species detected through visual identification or physical capture. The development of conservative standards for conducting and reporting eDNA analyses are necessary to publish eDNA detections in a public database. We plan to work with the invasive-species and eDNA communities to develop these standards and create a prototype platform for uploading eDNA results into the NAS database, thus allowing researchers and decision makers access to expanded species-presence data when they are making management decisions. We are seeking input from members of the CDI to facilitate our goal of synthesizing the various data and quality assurance requirements from a diverse group of eDNA data end-users.

Serving the U.S. Geological Survey's Geochronological Data

By Amy Gilmer and Leah Morgan (both USGS)

[Science Support Framework category: Data]

Geochronological data provide essential information necessary to address fundamental Earth-science questions. Understanding the timing of geologic processes and events as well as quantifying rates and time scales is key to geologic-mapping and assessments of mineral- and energy- resources and natural hazards. The U.S. Geological Survey's National Geochronological Database (NGDB) contains over 30,000 radiometric ages, including data generated by USGS and University laboratories, for rocks in the United States. The database was created in 1974 and was last fully updated in 1991. Whereas the data are currently accessible to the public through a web interface (<https://mrddata.usgs.gov/geochron/map-us.html>), these data have not been formally updated in more than 20 years. In the interim, the USGS has generated enormous amounts of geochronological data, many of which are difficult to access in a streamlined mechanism.

Age data need to be available and easily accessible to the public and to USGS scientists. For example, exploration geologists looking for critical minerals or energy resources need to know the ages of the rocks in order to identify potential targets for resources. Having these data readily accessible will also enhance decision making related to geologic hazards and enable efficient geologic mapping. There is recognition that USGS data need to be accessible beyond those within a specific project and externally outside of the USGS.

Gridding Airborne Radiometric Data Using the Stan Probabilistic Programming Language: An Example of Bayesian Statistical Modeling With RStudio and ArcGIS Pro

By Margaret Goldman (USGS)

[Science Support Framework category: Science Data Lifecycle—Analysis]

Airborne gamma-ray spectrometry (AGRS) measures the gamma rays that are emitted from naturally occurring radioactive isotopes found in rocks and soil, the most abundant of which are potassium (K40), uranium (U238), and thorium (Th232). Radiometric data can aid in the exploration for critical mineral resources, including deposits of barium, fluorine, titanium, beryllium, niobium, rare-earth elements, and uranium. There is also growing interest in using radiometric data to map soil

properties. The airborne radiometric data are an example of compositional data that are nonstationary (that is, the mean and the standard deviation vary spatially). It is therefore important to apply statistical techniques that account for both properties when creating maps. To this end, a Bayesian hierarchical model coded in the Stan probabilistic programming language (Ellefsen and others, 2020) is used to estimate spatial variations of the means and standard deviations for K40, eU238, and eTh232. The 2005 national airborne radiometric data were used to create new maps of these three radioactive isotopes in the conterminous United States (Goldman and Ellefsen, 2020).

SageDAT: Data and Tools to Support Collaborative Sagebrush-Ecosystem Conservation and Management

By Steven E. Hanser (USGS), Dell Long (USGS), Paul F. Steblein (USGS), Lief A. Wiechman (U.S. Fish and Wildlife Service), Karen L. Prentice (Bureau of Land Management), Ken E. Mayer (Western Association of Fish and Wildlife Agencies), Tim J. Kern (USGS), John C. Tull (U.S. Fish and Wildlife Service), and Michael E. Houts (Western Association of Fish and Wildlife Agencies)

[Science Support Framework category: Data Management]

Collaborative and science-informed management has been at the heart of the large-scale efforts to conserve the sagebrush ecosystem for greater sage-grouse (*Centrocercus urophasianus*) and more than 350 other species that rely on the sagebrush ecosystem. The development and use of geospatial data and decision-support tools to inform management of rangeland fires, restoration of sagebrush habitats, and conservation of the greater sage-grouse has resulted in exciting new opportunities, but the volume of data and tools has resulted in challenges for providers and users of this information. SageDAT is planned to be a web-based system that uses the latest technology to reduce barriers to data sharing and increase access to information through the development of a multiagency data catalog. For data providers, this effort is expected to increase communication and coordination on data management and provide tools to protect sensitive and (or) proprietary locations and information and thus alleviate past impediments to participation in large-scale planning efforts. For users, SageDAT is expected to provide access to a comprehensive list of datasets and decision-support tools through a web interface and improve mechanisms for increased communications and cooperation among Federal, state, and local agencies, tribes, nongovernmental organizations, universities, and industry to enhance long-term stewardship of the sagebrush ecosystem in 11 western States.

DataAtRisk.org and Its Community-Driven Data-Rescue Nomination Tool

By Sophie Hou (National Center for Atmospheric Research) and Reid Boehm (University of Houston)

[Science Support Framework category: Science Data Lifecycle—Preservation]

[DataAtRisk.org](https://dataatrisk.org/)'s (<https://dataatrisk.org/>) Data Nomination Tool facilitates community-driven rescue efforts for Earth- and environmental-science datasets. Particularly, the web-based tool connects people who can provide long-term data-stewardship support with those who need the assistance.

The concept is for the tool to have the following key functions:

- allowing a dataset to be submitted (or “nominated”) with a request for assistance through a web form;
- enabling someone who can help with the data-rescue activities to sign up and select the activities that meet their interests and capabilities; and
- sharing information about the [DataAtRisk.org](https://dataatrisk.org/) project's background and providing ways to contribute and (or) get involved.

If you need help or can provide help, or if you are simply interested in learning more about [DataAtRisk.org](https://dataatrisk.org/) and the current state of the prototype tool, we invite you to come by our poster and see a demo at the CDI Workshop.

The Data Nomination Tool was created and hosted by CloudBIRST (key contact: Joan Saez). [DataAtRisk.org](https://dataatrisk.org/)'s current members consist of individuals from Earth Science Information Partners (see ESIP Partners here: <https://www.esipfed.org/partners>), Johns Hopkins University Sheridan Libraries, and representatives from several University Research Libraries.

Measuring Sustainability of Community for Data-Integration Projects

By Leslie Hsu, Vivian Hutchison, and Madison Langseth (all from USGS)

[Science Support Framework category: Communities of Practice]

The Community for Data Integration (CDI) has been funding short-term data projects annually since 2010 and is interested in maximizing the sustainability and accessibility of the project outputs; however, there are no commonly accepted practices by which to measure the sustainability and accessibility of Earth-science-data project deliverables. Building on the work from

other disciplines that are addressing the sustainability of projects, we developed a framework for evaluating sustainability. Our framework describes seven sustainability influences and three ways of defining sustainability at the individual-, organization-, and community-levels. Using this framework, we evaluated outputs of projects funded by the USGS CDI. We found that the various outputs are widely accessible but not necessarily sustained or maintained. Projects with the highest number of sustainability influences we examined often became institutionalized and met a required need of the community. Even if proposed outputs were not delivered or sustained, knowledge of lessons learned could be passed along to others, building community capacity regarding a topic, which is another type of sustainability. We conclude by summarizing lessons about maximizing sustainability of projects for individuals applying for short-term funding and for organizations running programs that provide such funding.

USGS GGGSC Data Management and Spatial Studies (DMSS), Geophysical Inventory(s), Mobile Data Collection, and Automation

By Michaela R. Johnson, Margaret A. Goldman, Philip J. Brown, and Brian D. Rodriguez (all USGS)
[Science Support Framework category: Data Management]

The USGS Geology, Geophysics, and Geochemistry Science Center (GGGSC) Data Management and Spatial Studies (DMSS) project provides support for geospatial analyses; mobile field-data collection; management of geospatial collections including documentation; and distribution of all dataset types (geophysical, geochemistry, remote sensing (hyperspectral), and so on). GIS and data-management support, including hyperspectral and geophysical studies that improve capabilities and applications for investigating critical minerals, and data management and delivery support for the Earth Mapping Resources Initiative (Earth MRI), are provided. Vital and expensive geophysical-survey data that are only available internally (currently unpublished) will continue to be documented and published. New surveys will be tracked, documented, and disseminated to the public. Additionally, methods development, mobile GIS applications, automation, and data visualization are investigated. The project supports Mineral Resources Program mission-critical data themes focusing on geophysics and remote sensing.

Subsidence-Susceptibility Map for the Conterminous United States

By Jeanne Jones, Daniel Doctor, and Nathan Wood (all USGS)
[Science Support Framework category: Knowledge Management]

Sinkholes form in karst regions of the United States where voids in bedrock containing carbonate and evaporite minerals can be dissolved by water over time. Sinkholes present hazards to humans through subsidence and the drainage of contaminated surface-water runoff into groundwater, but also may play an important role as wetlands in some ecosystems. Sinkholes create instability in the foundations of buildings, roads, and other infrastructure, resulting in damage and, in some cases, loss of life. Geologists at the USGS have created maps showing the extent of bedrock that has the potential for karst erosion across the United States (Weary and Doctor, 2014); this project aims to use these maps along with techniques that use digital elevation data and a supercomputer to create the first nationwide digital dataset of sinkhole hot spots. This new map can be used by emergency managers, land-use planners, and public-works agencies.

Free Chocolate!!! Short Digital Object Identifier Usability Test Required

By Madison Langseth and Lisa Zolly (both USGS)
[Science Support Framework category: Publishing/Sharing]

Do you like free chocolate? Do you enjoy giving your opinion? Would you like to help improve a USGS data-management tool? If you answered “Yes!” to any of these questions, please stop by and participate in a short usability test for a single feature of the USGS Digital Object Identifier Tool.

A Generic Web Application to Visualize and Understand Movements of Tagged Animals

By Ben Letcher (USGS), and Jeff Walker (Walker Environmental Research LLC)
[Science Support Framework category: Applications]

The main goal of this project is to maximize the value of expensive animal-tagging data. We propose to do this by helping scientists understand patterns in their own tagging datasets and by helping scientists, funders, and agencies communicate tagging data to decision makers and to the general public. We also hope to contribute to and gain exposure for USGS data-visualization capability. Interactive visualization has emerged recently as a valuable tool for identifying patterns in complex datasets

that are typical of ecological-tagging studies. To make it easier and faster for users to gain access to interactive movement visualizations, we propose to develop the algorithms and the web-based software platform to allow users to upload their own data into a data-visualization file showing dynamic movements of tagged individuals across habitats. The overarching goal is to develop algorithms that are flexible enough to accommodate any animal-tagging data, and to provide a user-friendly interface that people will want to use with their own data. We are using six test cases ranging from one-dimensional stream networks to three-dimensional marine habitats to test the application and to receive feedback from data owners on application usability.

Building a Road Map for Making Data FAIR in the U.S. Geological Survey: A 2019 CDI Project

By Frances Lightsom (USGS)

[Science Support Framework category: Publishing/Sharing]

FAIR (findable, accessible, interoperable, reusable) is an international set of principles to promote multiple uses of data. Using the FAIR principles would improve the value of USGS data and tools by increasing their ease of use in downstream applications. Creating FAIR data would also promote USGS integrated science: readily available and compatible resources make projects easier because they can be integrated more efficiently.

A Road Map is an apt metaphor for this project's product. Our report will include information about the present status of the USGS application of FAIR principles; USGS goals for achieving compliance with the FAIR principles; obstacles and opportunities on the path to these goals; and a process for managing USGS progress that considers the inevitability of unexpected obstacles and opportunities.

Our project will involve the broad range of USGS staff who create, manage, and use scientific data. The central activity was a workshop in late summer 2019. You can help Make Data FAIR in the USGS by participating in the FAIR Workshop, completing a survey, or submitting a use case that will ensure that the Road Map meets the test of real data complexities.

Project team: Frances Lightsom, Vivian Hutchison, Natalie Latysh, Linda Debrewer, David Govoni, Wade Bishop, and Shelley Stall.

Improving a Classification System to Support FAIR Coastal and Marine Data

By Frances Lightsom (USGS)

[Science Support Framework category: Semantics]

Do you work with data that describe the physical habitats or biotic components of coastal or marine ecosystems? Are you wondering how to make your products more FAIR (findable, accessible, interoperable, reusable) by implementing metadata keywords and a data dictionary using a standard, interagency classification system? You are in luck: there is one such classification system, and it is called the Coastal and Marine Ecological Classification Standard (CMECS).

Are you currently working with CMECS but find it hard to apply in producing or integrating data? You're in luck: CMECS is a dynamic FGDC standard that is periodically updated in response to requests from data producers and users. This year, the CMECS team is collecting suggestions for changes to the standard.

Do you want to know more? Find us at the DataBlast, where we will have an online browser for exploring the hierarchy of CMECS categories and information on how to make the case for changes in CMECS. With luck, you'll also meet colleagues to help you define the change you need and to support your change request.

Coupling Hydrologic Models With Data Services in an Interoperable Modeling Framework

By Richard McDonald (USGS), Mark Piper (University of Colorado), Eric Hutton (University of Colorado), and Steven Markstrom (USGS)

[Science Support Framework category: Communities of Practice]

Integrated modeling is an important component of the U.S. Geological Survey's (USGS) Water Science Strategy. One approach is to develop a collaborative environment ("sandbox") to couple hydrologic and environmental-simulation models with data and analyses. We propose to leverage the existing "sandbox" developed by the Community Surface Dynamics Modeling System (CSDMS) team with several USGS models, including FaSTMECH (<https://i-ric.org/en/solvers/fastmech/>), PRMS6 (<https://www.usgs.gov/software/precipitation-runoff-modeling-system-prms>), and Modflow6 (<https://www.usgs.gov/mission-areas/water-resources/science/modflow-and-related-programs>), and the iRIC (International River Interface Cooperative, <https://i-ric.org/en/>) application Nays2Dflood (<https://i-ric.org/en/solvers/nays2dflood/>). The CSDMS Modeling Framework is used to wrap existing USGS models with Python interfaces. The models are then incorporated in the Python Modeling Toolkit (PyMT) to facilitate coupling. This project adds a new feature to the PyMT to provide model input from web-based data

services. Examples of model integrations include (1) coupling FaSTMECH to itself to extend a model domain, (2) coupling PRMS6 to Nays2DFlood to specify unengaged tributary flows in a flood-inundation simulation, and potentially (3) coupling Nays2DFlood to Modflow6 to improve flood-inundation predictions in areas where significant infiltration occurs during the flood duration. All software and examples developed through this project will be associated with an open-source license and made available through public repositories on GitHub (<https://github.com/csdms>).

Making Connections: The U.S. Geological Survey's National Digital Trails Network

By Elizabeth McCartney and Greg Matthews (both USGS)

[Science Support Framework category: Applications]

A major component of the Department of Interior's vision is to "Increase access to outdoor recreation opportunities for all Americans so that our people can be healthier, more fully enjoy the wonderful features of their Federal lands, and take advantage of hunting, fishing, and other outdoor recreation pursuits that are the roots of the conservation movement." The U.S. Geological Survey is advancing that vision with the launch of the National Digital Trails Network (NDT) project. The 2-year project consists of three major goals:

1. Develop a web-based geospatial-analysis tool to assist Federal land managers in identifying and prioritizing candidate trails for the connection of existing trails and trail networks.
2. Contribute to the creation of a robust national geospatial-trails dataset including, as a minimum, trails managed by key Federal agencies, including the U.S. Forest Service, National Park Service, U.S. Fish and Wildlife Service, and the Bureau of Land Management.
3. Develop a mobile responsive application that will assist trail stewards, land-management agencies, and members of the public, in the maintenance of trails information.

Join us for a discussion of the project and a demonstration of the Trail Routing, Analysis, and Information Linkage System (TRAILS) tool.

The Oceanographic Model and Data Portal: One Example of Integrating Interoperable Data

By Ellyn Montgomery (USGS)

[Science Support Framework category: Applications]

The sediment-transport group of the Woods Hole Coastal and Marine Science Center contracted Axiom Data Science to develop a portal (<https://cmgdata.usgsportals.net/>) as part of our Hurricane Sandy response. The portal was designed to provide easy browsing and display of information from already published resources: it's not just a metadata catalog, but also has extensive plotting capabilities and allows users to compare different datasets and see variations through time and space. The portal originally contained oceanographic time-series data and ocean-model grids; subsequently, other sources have been added, including tide gages, marsh units, and shoreline-change rates near Barnegat Bay, N.J. The underlying data are stored in CF- (Climate and Forecast) compliant netCDF (Network Common Data Form) files, in model-data grids served on THREDDS (Thematic Real-Time Environmental Distributed Data Services), or GIS layers that can be accessed on ScienceBase. We are currently experimenting with moving similar datasets published by other centers into the portal and hope it will become a widely used tool for discovering, displaying, understanding, and integrating various kinds of data.

Usage help is provided at <https://www.youtube.com/watch?v=SwXysu9z3yI> and <https://portal.aos.org/help/overview.html#data-views>.

ScienceBase as a Platform for Data Release: Workflow Automation

By Tamar Norkin and Ricardo McClees-Funinan (both USGS)

[Science Support Framework category: Science Data Lifecycle—Publishing/Sharing]

ScienceBase is a flexible platform that can facilitate data-integration tasks including, but not limited to, storing, managing, and distributing data; metadata display and maintenance; and creating web services.

Since 2015, the ScienceBase data-release team has provided a way to ensure quality, consistency, and the meaningful organization of USGS data release products through a standardized workflow and best practices. The workflow (see <https://www.sciencebase.gov/datarelease>) provides a clear, step-by-step path by which scientists can publish citable USGS data products in adherence to the requirements of the USGS data-release policies.

ScienceBase has recently built new features to automate key steps in the workflow and to manage the increase in data-release requests. These features include the ScienceBase Data Release Tool and a set of Jupyter Notebooks that the ScienceBase team uses to check and finalize data releases.

Advanced Computing Cooperative

By Courtney Owens, Janice Gordon, Richard Signell (all USGS)

[Science Support Framework category: Science Data Lifecycle—Processing]

The mission of the Advanced Computing Cooperative (ACC) is to help eliminate computational roadblocks by working across organizational boundaries to provide advanced computing capabilities to USGS researchers. The ACC is composed of volunteers who are focused on advancing the scientific-computing capabilities of the USGS and are from several USGS missions and centers.

The ACC has defined the following goals:

- Provide researchers with access to advanced computing resources, such as HPC, High-throughput Computing (HTC), and cloud-based advanced computing capabilities.
- Understand the Bureau requirements for advanced computing and define how the ACC can meet priority advanced-computing needs.
- Identify, prioritize, and address Bureau-level challenges and solutions to scientific computing problems.
- Provide consulting and training services to facilitate adoption and optimization of advanced-computing resources, workflows, and algorithms in USGS.
- Maintain and grow partnerships with external advanced-computing experts (for example, the Department of Energy (DOE), the Extreme Science and Engineering Discovery Environment (XSEDE) supported by the National Science Foundation, universities, and so on.)

The poster will define advanced computing and describe the role of the ACC, how the ACC helps researchers, and how you can get involved.

Reproducible Data Pipelines for Scientific Analyses

By Lindsay Platt (USGS)

[Science Support Framework category: Data Management]

Scientists conducting data analyses face a number of challenges in today's data-rich world, including efficiently dealing with the growing size and complexity of data, collaborating with team members who have various levels of data expertise and meeting requirements for disseminating models, data, and metadata. Whereas many members of the scientific community are embracing scripting languages to handle inputs and outputs of data, the increasing complexity of the analyses (for example, multiple data sources, access patterns, large data files) can strain the usefulness of basic scripting workflows. At the Water Mission Area Integrated Information Dissemination Division, the Data Science team has been exploring additional tools to achieve complete reproducibility, increase efficiency, bolster collaboration, and promote scalability in these complex situations. Features that have helped shape these robust workflows include modular and reproducible scripts, shared caching for intermediate data products, built-in fault tolerance when dealing with network transactions, and capturing dependencies among all inputs, processing steps, and outputs. Whereas we have implemented these features by using GNU-make and related R libraries [remake (FitzJohn, 2019), drake (Landau, 2018)], the concepts we've learned about these features can be applied to other languages. These practices can be taken one step further by integrating the tools with high-throughput computing to create powerful and manageable systems for data analysis.

Implementing a Grassland-Productivity Forecast Tool for the U.S. Southwest

By Sasha Reed (USGS), Bill Smith (University of Arizona), Brian Fuchs (National Drought Mitigation Center), Bill Parton (Colorado State University), Emile Elias (U.S. Department of Agriculture), and Brian Wardlow (Center for Advanced Land Management Information Technologies)

[Science Support Framework category: Applications]

Rangeland ecosystems are one of the largest types of providers of agro-ecological services in the United States and are particularly responsive to climate variability. The capacity to forecast rangeland-plant productivity for the upcoming year would greatly improve managers' ability to make decisions about stocking rates and locations, wildlife-forage needs, and fire-management plans. Here we describe the creation of a user-friendly, online tool that will provide forecasts of grassland and rangeland productivity for the southwestern United States and thus allow land managers, ranchers, scientists, and the general public to visualize and predict plant production for the upcoming season. We are integrating data from remote sensing, weather forecasting, and modeling techniques to build a system that provides updated plant-production forecasts every 2 weeks on a county-by-county scale. The financial and ecological security of the vast public lands and agroecosystems in the United States depends on our ability to manage resources in a dynamic world. The work described here focuses on the capacity of predictive science to address vulnerability, early warning, and decision support, and could have many uses, including decision making about wildlife, livestock, restoration, and fire.

GIS and Information Management Project—Putting Mineral Resources Program on the Map

By Carma San Juan, John Horton, Karen Lund, Stuart Giles, Jeff Mauk, and Steve Smith (all USGS)

[Science Support Framework category: Science Project Support]

The purpose of the GIS and Information Management Project is to support the geospatial component of the MRP. The project is organized by the Geology, Geophysics, and Geochemistry Science Center in Denver, Colo. The project provides a wide array of GIS support to MRP as well as assistance with publishing research results, new and legacy data, and web-enabled services. The project facilitates all aspects of the data lifecycle with the goal of effectively managing geospatial information from “cradle to grave.” GIS support to MRP projects encompasses acquiring, compiling, processing, analyzing, modeling, visualizing, documenting, reviewing, publishing, and archiving of geospatial products. The project develops and stewards national-scale data layers considered foundational to the MRP and conducts research to streamline workflows and to create innovative map products considered integral to the MRPs role in 21st-century science.

This poster highlights a few of the many MRP projects in which the GIS and Information Management Project has collaborated with researchers in the past 5 years. Geospatial products range in coverage from local to national-scale and span the breadth of MRP research. The authors note how published geospatial products often provide an important link to new research and program development in the MRP.

UAS Rad Cal: Open-Source Radiometric-Calibration Software

By Victoria Scholl, Calvin Kielas-Jensen, Dennis Helder, Josip Adams, Matthew Burgess, and Jeff Sloan (all USGS)

[Science Support Framework category: Applications]

USGS scientists are increasingly flying multispectral cameras on unmanned aircraft systems (UAS) to capture image data for their analyses. Radiometric calibration is an important image-processing step needed to assess changes in the landscape across space, time, and multiple sensors. This assessment involves converting raw Digital-Number (DN) pixel values to physical units such as reflectance. The USGS National UAS Project Office (NUPO) is developing a user-friendly open-source software option to support Department of the Interior scientists who do radiometric calibration of multispectral imagery. The USGS NUPO is currently testing and improving the Python-based software with multispectral imagery collected by using MicaSense RedEdge sensors and plans to support additional multispectral sensors in the future. This software can be used to provide empirical data with which to compare commercially available radiometric-calibration-software options to the USGS NUPO open-source-software option currently in development.

Pangeo: A Platform for Big-Data Geoscience on the Cloud

By Rich Signell (USGS)

[Science Support Framework category: Science Data Lifecycle—Analysis]

The Pangeo Project is a community-developed JupyterHub instance with a preconfigured open-source Python environment. It works on any Cloud (for example AWS, Microsoft Azure, Google Cloud, OpenStack) because it uses a Kubernetes cluster of Docker containers. To use the environment, a user needs only a browser and familiarity with Python, one of the most widely used programming languages in the scientific community. Parallel computation and memory management are done behind the scenes. The results of analysis can be explored by using interactive visualization tools. The result is a next-generation analysis and visualization platform for geoscience. Examples of processing Landsat imagery and simulated water levels during Hurricane Ike are demonstrated.

First Comprehensive Nonnative Species List for the United States, Segregated by Major Regions

By Annie Simpson (USGS) and Meghan C. Eyler (Natural Systems Analysts Inc.)

[Science Support Framework category: Data Management]

This nonnative species list consists of 11,344 unique taxa established in Alaska, Hawaii, the conterminous United States, or a combination of these regions. In all three regions, 157 taxa are established, and 1,166 authoritative sources were consulted to generate the list.

Our findings reinforce three common ideas: that tropical-island systems (in this case, Hawaii) are particularly vulnerable to biological invasions; that higher latitudes (in this case, Alaska) host fewer nonnative species but are not invulnerable to future invasions; and that species diversity in general decreases with increasing latitude.

Uses for the list include contributing to the measurement of Essential Biodiversity Variables for invasive-species monitoring, measuring gaps in coverage within species- occurrence databases, providing species references for early detection and rapid response, and assisting with prioritizing species incursions.

The nonnative species list was also added to expose nonnative occurrence records in BISON (<https://bison.usgs.gov>), an all-species mapping application that, as of May 2019, held more than 464 million native- and nonnative-species occurrence records (U.S. Geological Survey, 2015). By tagging BISON's nonnative occurrences, it was found that BISON contains about 18 million occurrence records for nonnative taxa, many of which were not labeled as nonnative by the data providers in the records.

Species-Occurrence Data for the Nation

By Annie Simpson, Elizabeth Sellers, and Elizabeth Martin (all USGS)

[Science Support Framework category: Applications]

BISON (Biodiversity Information Serving Our Nation, <https://bison.usgs.gov>), is a unique, web-based Federal mapping resource for species occurrence data in the United States, its Territories, its Marine Exclusive Economic Zones, and Canada (U.S. Geological Survey, 2015).

The size of the application is unprecedented: it includes more than 464 million records for most living species found in the United States and encompasses the efforts of more than a million professional and citizen scientists. Most of BISON's species-occurrence records are specific locations, not just county or state records.

If you wish to become a BISON data provider and make your data more broadly available through BISON (and optionally through the Global Biodiversity Information Facility, GBIF), please email bison@usgs.gov. As an application developed and supported by the Science Analytics and Synthesis Program of the USGS, the BISON project prioritizes acquisition and processing of datasets from USGS Science Centers, other agencies of the Department of the Interior, and other U.S. Federal and State agencies, followed by institutions, organizations, and individual researchers. BISON also seeks datasets with occurrence records specifically for invasive species or for pollinators.

Automated ScienceBase Data-Release Workflow: A Script to Update Metadata and Populate SB Pages

By Emily Sturdivant (USGS)

[Science Support Framework category: Publishing/Sharing]

This automation tool helps to create complex ScienceBase (SB) data releases. It is especially useful for populating data releases that are composed of many datasets, saving the author from the cumbersome process of creating many SB child pages and updating metadata files with the new child-page links.

The user starts with a populated SB landing page and a local-directory tree containing final data and metadata. The tool (a) creates SB child pages mimicking the input- directory structure, (b) updates all the XML (Extensible Markup Language) files with new SB links, (c) populates the SB pages with information parsed from the data/metadata, and (d) adds features such as browse graphics and bounding extents.

It uses Python ScienceBase Utilities (`sciencebasepy`) and was originally written for a specific dataset at the Woods Hole Coastal and Marine Science Center, but it has since been generalized to upload large datasets with FGDC metadata into ScienceBase. The tool is still in development and hosted on GitHub as `science-base-automation`. Depending on community interest, it could be integrated with existing tools.

Decoding NHD's VisibilityFilter Attribute: What Is It? Where Is It Available? and What Is the Accuracy?

By Hayley Thompson, Travis Landauer, Larry Stanislawski (all USGS)

[Science Support Framework category: Data]

The National Hydrography Dataset (NHD) includes a VisibilityFilter attribute that enables users to represent NHD vector features in the NHDPlus High Resolution (HR) at eight different map scales ranging from 1:24,000 up to 1:5,000,000. By using a feature-thinning model and workflow that approximates natural drainage-density patterns for the conterminous United States, each feature is assigned a VisibilityFilter value identifying an appropriate map scale and all larger scales for representing the features. This attribute is available for the NHDFlowline, NHDWaterbody, NHDArea, and NHDLine feature classes within NHDPlus HR. As the VisibilityFilter attribute is populated, data are available by download from <https://www.usgs.gov/core-science-systems/ngp/national-hydrography>.

As datasets with the VisibilityFilter attribute become available, ongoing research is being conducted to identify differences between datasets at the 1:100,000 map scale as defined by the VisibilityFilter in the NHDPlus HR data and the 1:100,000-scale NHD Medium Resolution data. The goal of this research is to identify patterns and areas of potential improvement in the methods used to calculate the VisibilityFilter attribute and to subsequently improve the accuracy in later versions of the VisibilityFilter.

Connecting Data to the National Hydrography Dataset

By Michael Tinker and Kevin McNinch (both USGS)

[Science Support Framework category: Communities of Practice]

The USGS creates and maintains the NHD, which portrays the surface water of the Nation. The NHD surface-water network provides a framework for linking data sources such as hydrologic observations, natural-resource surveys, and other water-related datasets.

The USGS is currently creating new methods by which hydrologic observations can be referenced to the NHD and shared as map services. For example, the USGS is creating a flexible data schema, called the Hydrography Referenced Data (HRD), that can be used for any kind of hydrologic observation. The USGS is also designing a Hydrography Referencing Tool (HRT), which will be a browser-based tool that allows linear referencing and flexible indexing of any kind of hydrologic observation to the NHD.

The HRT and HRD are both aligned with the vision of the National Hydrography Infrastructure and the Internet of Water. Supporters of these initiatives are collaborating on web tools and APIs for discovery and search of community generated HRD through a future web portal yet to be designed. This poster will show how the new HRD and HRT projects at the USGS fit within a collaborative community of holders of data related to the NHD.

ISO Made Easy: Content Specifications to Guide Metadata Authorship

By Dennis Walworth (USGS), Frances Lightsom (USGS), Tara Bell (USGS), Josh Bradley (U.S. Fish and Wildlife Service), Sophie Hou (National Center for Atmospheric Research), Andrew LaMotte (USGS), and Lisa Zolly (USGS)

[Science Support Framework category: Data Management]

The USGS will soon transition to the international metadata standards known collectively as ISO 19115. The open-ended nature of ISO provides much greater flexibility and vocabulary to describe research products; however, that flexibility means fewer constraints that can guide authors and ensure standardized, robust documentation across the Bureau. In FY18, CDI funded the ISO Content Specifications project to host a workshop to create a suite of modular specifications that would support the vast array of USGS data products. These modules are Basic (descriptive metadata common to all USGS data), Biological, Lineage, and Geospatial. Utilized together, these content-specification modules can create a compliant, descriptive metadata record. The team created an implied mapping to the ISO standard by crosswalking the content specifications to the ADIWg (Alaska Data Integration Working Group) mdJSON schema and thus enabling the creation of custom schemas and editor profiles that govern metadata-validation rules and customize the user interface in the mdEditor authoring tool. Users will see only the elements specified by the applied content-specification modules—streamlining metadata entry and guiding authors on content while ensuring robust, standardized USGS metadata. Perhaps most important, authors do not have to understand the intricacies of the ISO standard to write compliant metadata.

SHIRA Risk Mapper—Visualizing Multihazard Risk Exposure of DOI Lands, Populations, Assets, and Resources

By Nathan Wood, Jeanne Jones, Jason Sherba, Kevin Henry, and Peter Ng (all USGS)

[Science Support Framework category: Science Project Support]

The Department of Interior (DOI) protects and manages the natural resources and cultural heritage of the United States, provides information about those resources, and honors trust responsibilities or special commitments to American Indian, Alaska Native, and Insular and affiliated island communities. Significant numbers of people live, work, go to school, and recreate on lands managed or administered by the DOI.

Because of these responsibilities, the DOI Office of Emergency Management began a collaboration called the Strategic Hazard Identification and Risk Assessment (SHIRA) with the USGS on a Department of the Interior Resources Project to understand the risks posed by various hazards to its lands, facilities, people, revenues, and resources. Geospatial tools and data analytics are being used to identify variations in exposure to a wide array of hazards.

Key SHIRA deliverables include data analytics, web services, a web-based mapping application, a relative-threat matrix, and a data dashboard. These products can provide insights that support DOI officials' needs to make strategic decisions about how to mitigate risk, respond to incidents, allocate resources, and develop plans.

Summary of Workshop Outcomes

The 2019 Community for Data Integration (CDI) workshop provided an opportunity for colleagues to discuss shared areas of interest and to make new connections for future collaborations. There were three major outcomes of the workshop.

First, new projects, tools, and resources that are supported by the CDI (through its proposal process or through coordination of activities) were presented to the community. These presentations were accomplished through the DataBlast poster and demonstration session, the plenary lightning talks, and breakout sessions. An important role of the CDI is to inform its members about relevant resources, especially those supported by CDI funds or coordination, to its membership. By doing so, the CDI increases the awareness and use of project deliverables and the community expertise.

Second, USGS-wide resources that are useful for the large-scale integrated vision of USGS science were presented to the community. These presentations took place in the plenary talks, breakout sessions, and trainings, and included topics such as high-performance computing resources, the EarthMAP (Earth Monitoring, Analyses, and Prediction) vision of integrated predictive science, information about hosting applications in the USGS Cloud, analysis-ready data from the Landsat Program, and data- and software-management resources.

Third, workshop attendees identified several topics and questions that require future discussion and attention. These topics were documented through breakout-session key take-aways ([Appendix 3](#)) and through the audience-interaction software “sli.do” ([Appendix 4](#)). Notable topics were data visualization, usability, the importance of software in scientific research, collaboration with external partners, stakeholder needs, and artificial intelligence and machine-learning methods. The topics identified can be used to guide future CDI activities, events, and funding opportunities.

Acknowledgments

We would like to thank all the members of the Community for Data Integration (CDI), who made all of the CDI events and activities possible with their contributions of ideas and time. This includes the workshop attendees, who generated all of the content in this report; the CDI coordinators, who contributed ideas and advice from the beginning of the workshop-planning process; and the executive sponsors of the CDI, Kevin T. Gallagher, Tim Quinn, and Cheryl Morris. Workshop-committee volunteers included Madison Langseth, Leah Colasuonno, Tara Bell, Abby Benson, Paul Exter, Sophie Hou, Vivian Hutchison, Sue Kemp, Cassandra Ladino, Sophia Liu, Tamar Norkin, Marcia McNiff, Annie Simpson, and Nancy Sternberg. The author would like to thank Amanda Liford and Grace Donovan for assistance in compiling the proceedings information; and to Jacob Massey, Vivian Hutchison, and Daniel Wieferich for taking the photographs to document the workshop. Thank you to the reviewers Emily Brooks, Leah Colasuonno, Pete Ruhl, Chris Skinner, and Nancy Sternberg.

References Cited

- Chang, M.Y., Carlino, J.A., Barnes, C., Blodgett, D.L., Bock, A.R., Everette, A.L., Fernette, G.L., Flint, L.E., Gordon, J.M., Govoni, D.L., Hay, L.E., Henkel, H.S., Hines, M.K., Holl, S.L., Homer, C.G., Hutchison, V.B., Ignizio, D.A., Kern, T.J., Lightsom, F.L., Markstrom, S.L., O'Donnell, M.S., Schei, J.L., Schmid, L.A., Schoephoester, K.M., Schweitzer, P.N., Skagen, S.K., Sullivan, D.J., Talbert, C., and Warren, M.P., 2015, Community for Data Integration 2013 Annual Report: U.S. Geological Survey Open-File Report 2015–1005, 36 p., accessed March 07, 2018, at <https://doi.org/10.3133/ofr20151005>.
- Ellefsen, K.J., Goldman, M.A., and Van Gosen, B.S., 2020, User guide to the Bayesian modeling of non-stationary, univariate, spatial data using R language package BMNUS: U.S. Geological Survey Techniques and Methods, book 7, chap. C20, 27 p., accessed November 9, 2020, at <https://doi.org/10.3133/tm7C20>.
- 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 March 07, 2018, at <https://doi.org/10.3133/ofr20131265>.
- FitzJohn, R., 2019, remake—Make-like build management: R package, version 0.3.0, GitHub, accessed November 9, 2020, at <https://github.com/richfitz/remake>.
- Goldman, M.A., and Ellefsen, K.J., 2020, Bayesian modeling of NURE airborne radiometric data for the conterminous United States—Predictions and grids: U.S. Geological Survey data release, accessed November 9, 2020, at <https://doi.org/10.5066/P9YEAFHI>.
- 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 July 1, 2020, at <https://doi.org/10.3133/ofr20191123>.
- Landau, W.M., 2018, The drake R package—A pipeline toolkit for reproducibility and high-performance computing: *Journal of Open Source Software*, v. 3, no. 21, p. 550, <https://doi.org/10.21105/joss.00550>, accessed November 9, 2020, at <https://cran.r-project.org/web/packages/drake/index.html>.
- U.S. Geological Survey, 2015, Species occurrence data for the Nation—USGS Biodiversity Information Serving Our Nation (BISON), ver. 1.1, May 2019: U.S. Geological Survey General Information Product 160, 1 p., accessed October 9, 2020, at <https://doi.org/10.3133/gip160>.
- Weary, D.J., and Doctor, D.H., 2014, Karst in the United States—A digital map compilation and database: U.S. Geological Survey Open-File Report 2014–1156, 23 p., accessed November 9, 2020, at <https://dx.doi.org/10.3133/ofr20141156>.

Appendix 1.

Table 1.1 Agenda.

[All session leaders are with the U.S. Geological Survey except where specified. USGS, U.S. Geological Survey; CDI, Community for Data Integration; NCAR, National Center for Atmospheric Research; --, no data]

Time	Session title	Session leader(s)
Monday, June 3, 2019		
9:00 a.m.–4:30 p.m.	Science Gateways Community Institute session for CDI Funded Projects	Juliana Casavan and Claire Stirm, Science Gateways Community Institute
1:00 p.m.–4:00 p.m.	Intro to R Workshop	Lindsay Platt
Tuesday, June 4, 2019		
8:00 a.m.–8:30 a.m.	Registration	--
8:30 a.m.–10:00 a.m.	Welcome and Opening Remarks	Kevin T. Gallagher and Tim Quinn
	Turning your Data into Real Time Actionable Insights	Benjamin Tuttle, Arturo Inc.
10:00 a.m.–10:30 a.m.	Break	--
10:30 a.m.–12:00 p.m.	Lightning Talks	As submitted by attendees
	Partner Talks	Erin Robinson, Earth Science Information Partners Annie Burgess, Earth Science Information Partners Juliana Casavan and Claire Stirm, Science Gateways Community Institute
12:00 p.m.–1:30 p.m.	Lunch	--
1:30 p.m.–3:00 p.m.	Concurrent breakout sessions	
	Cloud Hosting Solutions: Service Offerings in the USGS Cloud	Jennifer Erxleben
	Preservation and Digitization of Physical Materials	Frances Lightsom
	Improving Collaboration Experiences between Data Providers and Curators	Sophie Hou, NCAR
	Data Management Workflow Show and Tell	Madison Langseth
3:00 p.m.–3:30 p.m.	Break	--
3:30 p.m.–5:00 p.m.	DataBlast Posters and Demos	As submitted by attendees
Wednesday, June 5, 2019		
8:30 a.m.–10:00 a.m.	Plenary Talks—Components of Integrated Science	
	National Hydrography Infrastructure	Sue Buto
	Risk Map	Nate Wood
	Advanced Research Computing	Janice Gordon
	Cloud Hosting Solutions Capabilities	Kimberly Scott
	ScienceBase	Drew Ignizio
	GeoPlatform	Tod Dabolt
10:00 a.m.–10:30 a.m.	Break	--
10:30 a.m.–12:00 p.m.	Plenary Talks—FAIR (Findable, Accessible, Interoperable, Reusable) Data and Models	
	FAIR 101	Shelley Stall, American Geophysical Union
	FAIR Benefits and Challenges in USGS	Jeanne Jones, Carma San Juan, Rebecca Scully
	Small Group Discussions of Successes and Challenges in USGS FAIR Practices	Wade Bishop, University of Tennessee, Knoxville
	FAIR and Data Fitness for Re-use Opportunities for Participation	
	Lightning Talks	As submitted by attendees
12:00 p.m.–1:30 p.m.	Lunch	--

Table 1.1 Agenda.—Continued

[All session leaders are with the U.S. Geological Survey except where specified. USGS, U.S. Geological Survey; CDI, Community for Data Integration; NCAR, National Center for Atmospheric Research; --, no data]

Time	Session title	Session leader(s)
1:30 p.m.–3:00 p.m.	Concurrent Breakout Sessions	
	Cloud Hosting Solutions 101: Demystifying the Cloud	Eric Larson
	Progress on Handling Large Data in the USGS and What's Ahead	Drew Ignizio
	Records Management and Data- Management Connections	Chris Bartlett
	Packaging Scientific Analyses as Software	Steve Aulenbach
	Want to know how to make your data applications delightful? Usability can help!	Madison Langseth and Sophie Hou
3:00 p.m.–3:30 p.m.	Break	--
3:30 p.m.–5:00 p.m.	Concurrent breakout sessions	
	Let's Talk Cloud	Michelle Guy
	Integrating Data and Models for Next-Generation Predictive Science	Ken Bagstad
	Data and Metadata, Review and Creation	Frances Lightsom
	Advanced Approaches to Data Management: Exploring Services and APIs	Drew Ignizio
	Science Gateways Community Institute Session: Building Your Value Proposition	Juliana Casavan and Claire Stirm
Thursday, June 6, 2019		
8:30 a.m.–10:00 a.m.	Plenary Talks—USGS Big Data to Smart Data	
	Ecological Forecasting: Making Smart Data from Big Data	Jake Weltzin
	Big Data from a Big Disaster: UAS Data Collection, Processing, and Dissemination During the 2018 Kilauea Eruption	Angie Diefenbach
	From Big Data to Smart Data in Remote Sensing	Pete Doucette
	Process-guided Deep Learning Predictions of Lake Water Temperature	Jake Zwart
	Evidence Based Policy and Department of Interior Data Governance	Tod Dabolt
10:00 a.m.–10:30 a.m.	Break	--
10:30 a.m.–12:00 p.m.	Concurrent breakout sessions	
	Let's Talk Cloud—Applications	Chris Soulard
	Content Specifications for ISO Metadata Standard	Dennis Walworth
	Software Release Questions Answered	Eric Martinez, Cassandra Ladino, Laura DeCicco
	Exploring the Landscape of Scientific Computing Tools for Large-Scale Analytics, Machine Learning, and Integrated Modeling	Janice Gordon and Jeff Falgout
12:00 p.m.–1:30 p.m.	Lunch	--
1:30 p.m.–3:00 p.m.	Concurrent breakout sessions	
	Data Management Working Group	Vivian Hutchison and Madison Langseth
	Citizen-Centered Innovation Working Group	Sophia Liu
	Software Development Cluster	Michelle Guy
	Tools You Can Use: Visualization—Community for Data Integration (CDI) Risk Map, TerriaMap, Tableau	Kevin Henry and Dionne Zoanni
3:00 p.m.–3:30 p.m.	Break	--
3:30 p.m.–5:00 p.m.	Award ceremony	
	Plenary discussion: The Future of the Community for Data Integration (CDI)	
	Closing Discussion on next steps	
Friday, June 7, 2019		
8:00 a.m.–4:00 p.m.	ISO Metadata Content Specification Workshop	Dennis Walworth, Frances Lightsom, Lisa Zolly
8:00 a.m.–12:00 p.m.	USGS Software-Release Practicum	Lance Everette
9:00 a.m.–12:00 p.m.	High-Performance Computing	Janice Gordon

Appendix 2.

Table 2.1 Attendees.

List of registered conference attendees. Personal email addresses have been removed. (CIRES: Cooperative Institute for Research in Environmental Sciences, DOI: Department of the Interior, NCAR: National Center for Atmospheric Research, NOAA: National Oceanographic and Atmospheric Administration, UNAVCO: official name (no longer an acronym), USGS: U.S. Geological Survey)

First Name	Last Name	Email	Affiliation
Sinan	Abood	sinanayadabood@fs.fed.us	U.S. Department of Agriculture
Seth	Ackerman	sackerman@usgs.gov	USGS
Joe	Adams	jdadams@usgs.gov	USGS
Chuck	Anderson	charles.anderson@noaa.gov	CU-Boulder CIRES/NOAA
Caitlin	Andrews	candrews@usgs.gov	USGS
Christy-Ann	Archuleta	carchule@usgs.gov	USGS
Matthew	Arsenault	marsenault@usgs.gov	USGS
Steve	Aulenbach	saulenbach@usgs.gov	USGS
Neil	Baertlein	nbaertlein@usgs.gov	USGS
Ken	Bagstad	kjbagstad@usgs.gov	USGS
Aparna	Bamzai-Dodson	abamzai@usgs.gov	USGS
Joe	Bard	jbard@usgs.gov	USGS
Abhijeeth	Baregal	abaregal@contractor.usgs.gov	USGS
Genevieve	Barron	gbarron@usgs.gov	USGS
Chris	Bartlett	cbartlett@usgs.gov	USGS
Jen	Bayer	jbayer@usgs.gov	USGS
Daniel	Beckman	dbeckman@usgs.gov	USGS
Tara	Bell	tbell@usgs.gov	USGS
George	Bennett	georbenn@usgs.gov	USGS
Abby	Benson	albenison@usgs.gov	USGS
Janelda	Biagas	biagasj@usgs.gov	USGS
Wade	Bishop	wade.bishop@utk.edu	University of Tennessee
Hannah	Boggs	hboggs@usgs.gov	USGS
John	Brakebill	jwbrakeb@usgs.gov	USGS
Sky	Bristol	sbristol@usgs.gov	USGS
Mary	Bucknell	mbucknell@usgs.gov	USGS
Annie	Burgess	annieburgess@esipfed.org	Earth Science Information Partners
Thomas	Burley	teburley@usgs.gov	USGS
Alan	Butler	rabutler@usbr.gov	Bureau of Reclamation
Kenna	Butler	kebutler@usgs.gov	USGS
Sue	Buto	sbuto@usgs.gov	USGS
Joe	Carroll	jcarroll@usgs.gov	USGS
Juliana	Casavan	--	Science Gateways Community Institute
Susan	Cochran	scochran@usgs.gov	USGS
Alexandra	Cohen	--	UNAVCO

Table 2.1 Attendees.—Continued

List of registered conference attendees. Personal email addresses have been removed. (CIRES: Cooperative Institute for Research in Environmental Sciences, DOI: Department of the Interior, NCAR: National Center for Atmospheric Research, NOAA: National Oceanographic and Atmospheric Administration, UNAVCO: official name (no longer an acronym), USGS: U.S. Geological Survey)

First Name	Last Name	Email	Affiliation
Leah	Colasuonno	lcolasuonno@usgs.gov	USGS
Bill	Condon	wcondon@usgs.gov	USGS
Margo D.	Corum	mcorum@usgs.gov	USGS
Joshua	Coyan	jcoyan@usgs.gov	USGS
Tom	Cram	tcram@ucar.edu	NCAR
VeeAnn	Cross	vatnipp@usgs.gov	USGS
Tod	Dabolt	thomas_dabolt@ios.doi.gov	Department of the Interior
Sofia	Dabrowski	sdabrowski@contractor.usgs.gov	USGS
Evan	Dailey	edailey@contractor.usgs.gov	USGS
Brian	Davis	bdavis@tableau.com	Tableau
Matthew	Davis	mdavis@contractor.usgs.gov	USGS
Linda	Debrewer	lmdebrew@usgs.gov	USGS
Jon	Dewitz	dewitz@usgs.gov	USGS
Angie	Diefenbach	adiefenbach@usgs.gov	USGS
Peter	Doucette	pdoucette@usgs.gov	USGS
Blake	Draper	bdraper@usgs.gov	USGS
Jason	Duke	jason_duke@fws.gov	U.S. Fish & Wildlife Service
Sara	Eldridge	seldridge@usgs.gov	USGS
Kenneth	Elsner	kenneth_elsner@fws.gov	U.S. Fish & Wildlife Service
Frank	Engel	fengel@usgs.gov	USGS
Kyle	Enns	kenns@usgs.gov	USGS
Jennifer	Erleben	jerleben@usgs.gov	USGS
Nick	Estes	njestes@usgs.gov	USGS
Lance	Everette	everettel@usgs.gov	USGS
Paul	Exter	peexter@usgs.gov	USGS
Jeff	Falgout	jfalgout@usgs.gov	USGS
Jeremy	Fee	jmfee@usgs.gov	USGS
Ben	Feinstein	bfeinstein@usgs.gov	USGS
Jason	Ferrante	jferrante@usgs.gov	USGS
Emily	Fort	efort@usgs.gov	USGS
Aaron	Fox	afox@usgs.gov	USGS
Mike	Frame	mike_frame@usgs.gov	USGS
Kevin	Frye	kfrye@tableau.com	Tableau
Jackson	Galloway	--	USGS
Christopher	Garrity	cgarrity@usgs.gov	USGS
Amy	Gilmer	agilmer@usgs.gov	USGS

Table 2.1 Attendees.—Continued

List of registered conference attendees. Personal email addresses have been removed. (CIRES: Cooperative Institute for Research in Environmental Sciences, DOI: Department of the Interior, NCAR: National Center for Atmospheric Research, NOAA: National Oceanographic and Atmospheric Administration, UNAVCO: official name (no longer an acronym), USGS: U.S. Geological Survey)

First Name	Last Name	Email	Affiliation
Maggie	Goldman	mgoldman@usgs.gov	USGS
Janice	Gordon	janicegordon@usgs.gov	USGS
Gregory	Gunther	ggunther@usgs.gov	USGS
Michelle	Guy	mguy@usgs.gov	USGS
Steve	Hanser	shanser@usgs.gov	USGS
Kaely	Harris	--	UNAVCO
Travis	Harrison	tharrison@usgs.gov	USGS
Alex	Headman	AHeadman@usgs.gov	USGS
Mike	Hearne	mhearne@usgs.gov	USGS
Kevin	Henry	khenry@usgs.gov	USGS
Robin	Hilderman	--	USGS
Mary	Hill	mchill@ku.edu	University of Kansas
Donn	Holmes	daholmes@usgs.gov	USGS
Marla	Hood	mhood@usgs.gov	USGS
Cheryl	Horton	cahorton@usgs.gov	USGS
Sophie	Hou	hou@ucar.edu	NCAR
Harry	House	hrhouse@usgs.gov	USGS
Daniel	Howard	dhoward3@nd.edu	University of Notre Dame
Leslie	Hsu	lhsu@usgs.gov	USGS
Marc	Hunter	mahunter@usgs.gov	USGS
Maggie	Hunter	mhunter@usgs.gov	USGS
Liz	Huselid	ehuselid@usgs.gov	USGS
Vivian	Hutchison	vhutchison@usgs.gov	USGS
Drew	Ignizio	dignizio@usgs.gov	USGS
Vaughn	Ihlen	vihlen@usgs.gov	USGS
Karen	Jenni	kjenni@usgs.gov	USGS
Mikki	Johnson	mrjohns@usgs.gov	USGS
Brian	Johnson	brian.johnson-1@colorado.edu	USGS
Jeanne	Jones	jmjones@usgs.gov	USGS
Kimberly	Jones	kjones@usgs.gov	USGS
Eric	Jones	esjones@usgs.gov	USGS
Max	Joseph	maxwell.b.joseph@colorado.edu	USGS
Alex	Kaufman	akaufman@usgs.gov	USGS
Daniel	Kelly	--	USGS
Sue	Kemp	skemp@usgs.gov	USGS
Tim	Kern	kernt@usgs.gov	USGS

Table 2.1 Attendees.—Continued

List of registered conference attendees. Personal email addresses have been removed. (CIRES: Cooperative Institute for Research in Environmental Sciences, DOI: Department of the Interior, NCAR: National Center for Atmospheric Research, NOAA: National Oceanographic and Atmospheric Administration, UNAVCO: official name (no longer an acronym), USGS: U.S. Geological Survey)

First Name	Last Name	Email	Affiliation
Simon	Kingston	simon_kingston@nps.gov	National Park Service
Kristi	Kline	kkline@usgs.gov	USGS
Rudy	Klucik	rudy.klucik@noaa.gov	CU-Boulder CIRES/NOAA
Ben	Knott	bknot@esri.com	Esri
Carl	Krupitzer	carl@thinglogix.com	ThingLogix
Cassandra	Ladino	ccladino@usgs.gov	USGS
Andy	LaMotte	alamotte@usgs.gov	USGS
Madison	Langseth	mlangseth@usgs.gov	USGS
Eric	Larson	edlarson@usgs.gov	USGS
Natalie	Latysh	nlatysh@usgs.gov	USGS
Gary	Latzke	gdlatzke@usgs.gov	USGS
Neda	Ledoux	nledoux@contractor.usgs.gov	Cherokee Nation Technologies-Contractor to DOI-USGS
Ben	Letcher	bletcher@usgs.gov	USGS
Chris	Lett	chris_lett@fws.gov	USGS
Anna	Liao	anna.liao@noaa.gov	CU-Boulder CIRES/NOAA
Amanda	Liford	amanlifo@vols.utk.edu	University of Tennessee
Frances	Lightsom	flightsom@usgs.gov	USGS
Sophia B	Liu	sophialiu@usgs.gov	USGS
Dell	Long	jllong@usgs.gov	USGS
Ryan	Longhenry	rlonghenry@usgs.gov	USGS
Ximena	Lopez Zieher	zlopez@agro.uba.ar	University of Buenos Aires, Argentina
CJ	Loria	cjloria@usgs.gov	USGS
Kris	Ludwig	kaludwig@usgs.gov	USGS
Tim	Mancuso	tmancuso@usgs.gov	USGS
Heather	Manley	hbattlesmanley@contractor.usgs.gov	Cherokee Nation Technologies-Contractor to DOI-USGS
William	Marken	wmarken@usgs.gov	USGS
Eric	Martinez	emartinez@usgs.gov	USGS
Sam	Martinez	scmartinez@usgs.gov	USGS
Derek	Masaki	dmasaki@usgs.gov	USGS
Robert	Matthias	rmatthias@usgs.gov	USGS
Matthew	Mayernik	mayernik@ucar.edu	NCAR
Elizabeth	McCartney	emccartney@usgs.gov	USGS
Ricardo	McClees-Funinan	rmcclees-funinan@usgs.gov	USGS
John	McCoy	mccoyj@usgs.gov	USGS
Richard	McDonald	rmcd@usgs.gov	USGS

Table 2.1 Attendees.—Continued

List of registered conference attendees. Personal email addresses have been removed. (CIRES: Cooperative Institute for Research in Environmental Sciences, DOI: Department of the Interior, NCAR: National Center for Atmospheric Research, NOAA: National Oceanographic and Atmospheric Administration, UNAVCO: official name (no longer an acronym), USGS: U.S. Geological Survey)

First Name	Last Name	Email	Affiliation
Joe	McInerney	joemci@ucar.edu	NCAR
Marcia	McNiff	mmcniff@usgs.gov	USGS
Kevin	McNinch	klmeninch@usgs.gov	USGS
Mckenzie	Metzger	mrmetzger@contractor.usgs.gov	USGS
Rob	Miller	rfmiller@usgs.gov	USGS
Heather	Miller	hmillier@usgs.gov	USGS
Mark	Miller	mpmiller@usgs.gov	USGS
Burke	Minsley	bminsley@usgs.gov	USGS
Ben	Mirus	bbmirus@usgs.gov	USGS
Ellyn	Montgomery	emontgomery@usgs.gov	USGS
Leah	Morgan	lemorgan@usgs.gov	USGS
Cheryl	Morris	cmorris@usgs.gov	USGS
Jim	Nagode	jbnagode@usbr.gov	Bureau of Reclamation
Matthew	Neilson	mneilson@usgs.gov	USGS
Jeremy	Newson	jknewson@usgs.gov	USGS
Lily	Niknami	lniknami@usgs.gov	USGS
Tamar	Norkin	tnorkin@usgs.gov	USGS
Ray	Obuch	obuch@usgs.gov	USGS
Allison	Odell	aodell@usbr.gov	Bureau of Reclamation
Sheryn	Olson	sherynolson@usgs.gov	USGS
Shann	O'Neill	--	USGS
Jerry	Ornelas	jxornelas@usgs.gov	USGS
Cory	Overton	covertton@usgs.gov	USGS
Courtney	Owens	clowens@usgs.gov	USGS
Andy	Park	apark@usgs.gov	USGS
Jeff	Pearson	jpearson@esri.com	Esri
Scott	Peckham	scott.peckham@colorado.edu	CU Boulder
Emily	Perkins	eperkins@usgs.gov	USGS
Roy	Petrakis	rpetrakis@usgs.gov	USGS
Mark	Piper	mark.piper@colorado.edu	University of Colorado
Lindsay	Platt	lpplatt@usgs.gov	USGS
Lindsay	Powers	lpowers@usgs.gov	USGS
Lauren	Privette	lprivette@usgs.gov	USGS
Tim	Quinn	tsquinn@usgs.gov	USGS
Rob	Rastovich	rob@thinglogix.com	Thinglogix
Sasha	Reed	screed@usgs.gov	USGS

Table 2.1 Attendees.—Continued

List of registered conference attendees. Personal email addresses have been removed. (CIRES: Cooperative Institute for Research in Environmental Sciences, DOI: Department of the Interior, NCAR: National Center for Atmospheric Research, NOAA: National Oceanographic and Atmospheric Administration, UNAVCO: official name (no longer an acronym), USGS: U.S. Geological Survey)

First Name	Last Name	Email	Affiliation
Mark	Reidy	mreidy@usgs.gov	USGS
Jodi	Riegle	jlriegle@usgs.gov	USGS
Sally	Roberts	sroberts@usgs.gov	USGS
Erin	Robinson	erinrobinson@esipfed.org	Earth Science Information Partners
George	Rolston	grolston@contractor.usgs.gov	USGS
Sarah	Ryker	sryker@usgs.gov	USGS
Carma	San Juan	csanjuan@usgs.gov	USGS
Erika	Sanchez-Chopitea	esanchez-chopitea@usgs.gov	USGS
Mark	Sandstrom	sandstro@usgs.gov	USGS
Victoria	Scholl	vscholl@usgs.gov	USGS
George	Schrader	george_schrader@fws.gov	U.S. Fish and Wildlife Service
Heather	Schreppel	hschreppel@contractor.usgs.gov	USGS
Doug	Schuster	schuster@ucar.edu	NCAR
Peter	Schweitzer	pschweitzer@usgs.gov	USGS
Kimberly	Scott	kvscott@usgs.gov	USGS
Rebecca	Scully	rscully@usgs.gov	USGS
Gabriel	Senay	senay@usgs.gov	USGS
Robert	Shepherd	rshepherd@usgs.gov	USGS
Jason	Sherba	jsherba@usgs.gov	USGS
Lauren	Sherson	lsherson@usgs.gov	USGS
Rich	Signell	rsignell@usgs.gov	USGS
James	Sill	jsill@esri.com	Esri
Julie	Simon	jlsimon@usgs.gov	USGS
Annie	Simpson	asimpson@usgs.gov	USGS
Chris	Skinner	cskinner@usgs.gov	USGS
Chris	Slater	chris.slater@noaa.gov	CU-Boulder CIRES/NOAA
Deborah	Smith	deborahsmith@usgs.gov	USGS
Tom	Sohre	tsohre@usgs.gov	USGS
Chris	Soulard	csoulard@usgs.gov	USGS
Deb	Spahr	dsspahr@usgs.gov	USGS
Shelley	Stall	sshall@agu.org	American Geophysical Union
Aaron	Stephenson	astephenson@usgs.gov	USGS
Nancy	Sternberg	nsternberg@usgs.gov	USGS
Jens	Stevens	jtstevens@usgs.gov	USGS
Rachel	Stevenson	rstevenson@usgs.gov	USGS
Claire	Stirm	cfrist@purdue.edu	Science Gateways Community Institute

Table 2.1 Attendees.—Continued

List of registered conference attendees. Personal email addresses have been removed. (CIRES: Cooperative Institute for Research in Environmental Sciences, DOI: Department of the Interior, NCAR: National Center for Atmospheric Research, NOAA: National Oceanographic and Atmospheric Administration, UNAVCO: official name (no longer an acronym), USGS: U.S. Geological Survey)

First Name	Last Name	Email	Affiliation
Emily	Sturdivant	esturdivant@usgs.gov	USGS
Colin	Talbert	talbertc@usgs.gov	USGS
Hayley	Thompson	--	USGS
Michael	Tinker	mdtinker@usgs.gov	USGS
Cody	Trenholm	--	UNAVCO
Matt	Tricomi	mtricomi@ios.doi.gov	Xentity
Dorothy	Trujillo	dtrujillo@usgs.gov	USGS
Gregory	Tucker	gtucker@colorado.edu	University of Colorado, Boulder
Benjamin	Tuttle	ben@arturo.ai	Arturo
Shayne	Urbanowski	surbanowski@usgs.gov	USGS
Dalia	Varanka	dvaranka@usgs.gov	USGS
Roland	Viger	rviger@usgs.gov	USGS
Miguel	Villarreal	mvillarreal@usgs.gov	USGS
Hans	Vraga	hvraga@usgs.gov	USGS
Jessica	Walker	jjwalker@usgs.gov	USGS
Jennifer	Walter	jlwalter@usgs.gov	USGS
Dennis	Walworth	dwalworth@usgs.gov	USGS
Tristan	Wellman	twelman@usgs.gov	USGS
Jake	Weltzin	jweltzin@usgs.gov	USGS
Rob	Wertz	rwertz@usgs.gov	USGS
Michael	Wieczorek	mewieczo@usgs.gov	USGS
Daniel	Wieferich	dwieferich@usgs.gov	USGS
Paul	Wiese	pmwiese@usgs.gov	USGS
Steve	Williams	sfw@ucar.edu	NCAR
Brad	Williams	bradwilliams@usgs.gov	USGS
Mark	Wiltermuth	mwiltermuth@usgs.gov	USGS
John	Wolf	jwolf@usgs.gov	USGS
Kevin	Wood	wood@usgs.gov	USGS
Nate	Wood	nwood@usgs.gov	USGS
Cheryl	Woodall	cwoodall@usgs.gov	USGS
Lynn	Yarmey	yarmel@rpi.edu	Research Data Alliance
Dionne	Zoanni	dzoanni@usgs.gov	USGS
Lisa	Zolly	lisa_zolly@usgs.gov	USGS
Jacob	Zwart	jzwart@usgs.gov	USGS

Appendix 3. Key Take-aways

Session leaders were asked to contribute key take-aways from their session. These take-aways are organized below in alphabetical order of the session title. The session title and session leader's last name are listed to identify the session.

Advanced Approaches to Data Management: Exploring Services and APIs (Ignizio)

1. There is a need to pull data from core systems, such as BASIS+, Active Directory, IPDS (the USGS Information Product Data System), ScienceBase, GitLab, the USGS Publications Warehouse, and others, to populate center-level tracking systems. These USGS science center systems usually are customized to meet center needs.
2. API (application programming interface) access is critical for systems. Read-only APIs will work for most, but some might be valuable to write to as well.
3. Sharing Python services, automated jobs, and other tools across the USGS would be beneficial to many users.
4. Embrace creativity to solve tough problems.

Collaboration Area—Citizen-Centered Innovation (Liu)

1. The Citizen-Centered Innovation Collaboration Area is producing a guidance document to address the growing interest in crowdsourcing, citizen science, and prize competitions.
2. The Collaboration Area shares case studies and best practices so that members can learn from successful examples and understand what happened when these techniques fail.
3. The collaboration area is developing a Department of the Interior Generic Information Collection Request (ICR) for Open Innovation. The ICR will help to reduce the policy barriers by enabling and accelerating innovation.

Collaboration Area—Data Management (Hutchison)

1. The Data Management collaboration area accomplished a large amount of networking through “Speed Data-ing.” The session activity identified 30 people with similar challenges, 23 people with similar strengths, and 22 instances where one person's challenge was matched to another person's strength.
2. The session discussion identified 10 topical themes to be addressed in the data management working group, including communicating about data management, automating processes, and exploring creative metadata approaches.
3. The next step is to plan Data Management Working Group monthly agendas around these topics.

Collaboration Area—Software-Development Cluster (Guy)

1. The USGS mission is our biggest strength for attracting software-development talent, but hiring could be better coordinated to support career development.
2. There are three tiers of support with various strengths and weaknesses: the USGS-wide Office of Enterprise Information, distinct Centers and Programs at the USGS, and communities of practice like those at the CDI.
3. An agency-specific tool to find people, skills, code reviewers, and existing software would improve collaboration.

Cloud Hosting Solutions 101: Demystifying the USGS Cloud (Larson)

1. To get started in the USGS Cloud and learn about current capabilities, contact the CHS team to make sure you can distinguish between common myths and facts.
2. CHS offers a variety of services, including Sandbox, Managed Services, and Custom Environments, to meet your specific need.
3. Additional information can be found at <https://support.chs.usgs.gov/> and on the USGS internal site of the Office of Enterprise Information (CHS) site. General CHS inquiries may be sent to cloudservices@usgs.gov.

Cloud Hosting Solutions: Service Offerings in the USGS Cloud (Erxleben)

1. To support more complex modeling capabilities and the EarthMap vision, CHS continues to develop and expand the ecosystem of Cloud-based services and tools that will allow scientists to acquire, process, analyze, model, store, share, and disseminate their data in a common environment.
2. CHS continues to work with the Grand Canyon Monitoring and Research Center to develop innovative technologies to advance monitoring and prediction of processes in the Colorado River Basin through the Cloud Sensor Processing Framework.
3. Next steps include exploring cloud-based artificial intelligence and machine learning capabilities and continuing to improve communication.

Content Specifications for ISO Metadata Standard (Walworth)

1. Use smart metadata and documentation. This means standardized, robust documentation throughout the USGS. Help authors focus on content rather than struggle with XML standards.
2. Join the community that is working to transition to the ISO metadata standard. Our path is defined by the science data community, and there is the opportunity to define USGS documentation requirements.
3. The group is thinking about governance related to the use of the ISO metadata standard at USGS.

Data and Metadata, Review and Creation (Lightsom)

1. We need more metadata trainings for new people and supervisors.
2. We need to tackle how we are translating data and metadata into ISO metadata (International Organization for Standards, whose metadata standards are known collectively as ISO 19115).
3. Stewardship of data and metadata is important.

Data-Management-Workflow Show and Tell (Langseth)

1. It would be beneficial to collaborate and share more across the USGS for implementing data management planning and other data management activities.
2. It would be beneficial to integrate a records-management approach with data-management processes (for example, a single Data and Records Management Plan).
3. Next Steps: Share what data management staffing looks like at different centers, and reach out to someone from the Coastal and Marine Geology Program (CMGP) to talk about Compass for data management planning, and whether something similar could be adopted by other centers.

Exploring the Landscape of Scientific-Computing Tools for Large-scale Analytics, Machine Learning, and Integrated Modeling (Gordon)

1. There is open communication about future computing needs through the Community for Data Integration and working groups in the Advanced Computing Consortium.
2. Current working groups are working towards operational HPC.
3. To work efficiently in this field, keep up on training, especially on new machines such as Tallgrass and Denali.

Improving Collaboration Experiences Between Data Providers and Curators (Hou)

1. Communicate about benefits, and agree on goals.
2. Find effective balance points between independent and collaborative activities.
3. Leverage user-friendly data management tools and data services.

Integrating Data and Models for Next-Generation Predictive Science (Bagstad)

1. Of the integrated modeling frameworks and semantics presented in this session, we see convergence and similarities. When are the times in the scientific- integration lifecycle to promote the convergence of integrated modeling frameworks, and how does one promote the convergence of disparate systems? Alternatively, how do we recognize the times when systems are in experimental stages and it's wiser to "let a thousand flowers bloom"?
2. Convergence to a unified modeling framework may help if the funding of any one development group is uncertain. Modelers may hesitate to buy into a framework if its developers have been defunded.
3. Semantics will be necessary for machine readability and actionability for data and model integration. The semantic foundations for representing interdisciplinary science could be reused to support USGS efforts but would be somewhat different from discipline-specific semantics linked through thesauri.

Let's Talk Cloud (Guy)

1. Cloud is dynamic and flexible.
2. Continue working with CHS (Cloud Hosting Solutions) to explore solutions.
3. Learn from and with other CHS customers.
4. Once you are in, many possibilities open up, so keep trying!

Let's Talk Cloud–Applications (Soulard)

1. The benefits of Google Earth Engine are great and necessary for cutting-edge remote sensing research and development.
2. Security concerns have led to our current (fiscal year 2019) situation where access to Google Earth Engine is restricted.
3. USGS leadership is hopeful that we can find a resolution in the future to help scientists do their research effectively.

Packaging Scientific Analyses As Software (Wiltermuth)

1. Packaging scientific analyses as software increases reproducibility, replicability, and transparency.
2. Packaging scientific analyses as software improves the simplicity of providing all research artifacts for potential interactive delivery of information. We need to begin exploring where USGS can lead in this area—for example, an option for new USGS publication type.
3. Including a record of stakeholder engagement and input into analysis methods is essential for tracking the provenance of project ideas and purpose.
4. We discussed diverse approaches to packaging analyses in a wide range of applications within the USGS; however, more discussion is needed to guide future directions, and provide recommendations for people in the beginning stages.

Preservation and Digitization of Physical Materials (Lightsom)

1. You can do a lot with a little. We all have few resources; we just need to be creative. One way is by using students and volunteers through the STEP-UP program (Secondary Transition to Employment Program–USGS Partnership).
2. Don't be afraid to try out new software for managing collections.
3. There is a records-management challenge related to collections in the USGS, and thus we have the opportunity to continue to refine and improve our processes for the benefit of future data.
4. There is a lot of expertise in USGS regarding physical collections.
5. It's ok to find a stopping point for a preservation project, otherwise, these activities could go on and on.

Progress on Handling Large Data in the USGS and What's Ahead (Ignizio)

1. The Black Pearl storage device is a new option for Advanced Research Computing (ARC) users, USGS users in general, and users doing ScienceBase submission. This capability can be used by client or Command Line Interface (CLI) and is now self-serve with Active Directory credentials.
2. Globus large-file transfer process will soon be available to more users and is becoming more resilient to time-outs and other issues.
3. The National Center for Atmospheric Research (NCAR) has a lot of experience with different access approaches (for example, services) for netCDF data. It would be beneficial for USGS to work with NCAR and learn from their approaches.
4. Cloud-based files may be best served in some newer formats (for example, Zarr). These formats are much faster, but require conversion, and have a learning curve.

Records Management and Data Management Connections (Bartlett)

1. There are opportunities to connect the dots between records management and data management. Data are records that must be preserved in the context of other records management activities.
2. Organization, preservation, and storage are three key records management aspects.
3. Records management is a collaborative process between the Records Management Program and the USGS, for example, through standard file creation, records-schedule development, preservation, storage requirements, and so much more.

Science Gateways Community Institute Communication Session (Casavan)

1. A value proposition provides a clear understanding of the unique value your project delivers to your users or stakeholders.
2. You can craft one with the following template: [MY PROJECT] Will help [WHO?] Do [WHAT?] By [HOW?]
3. Once the value proposition is crafted, it still needs to be communicated before it is useful!

Software Release Questions Answered (Martinez)

1. A new instructional memorandum (IM) from the Office of Science Quality and Integrity, IM OSQI 2019–01, is coming soon. The IM delegates authority and discretion for software release to center directors. (Note: This IM has since been released at <https://www.usgs.gov/about/organization/science-support/survey-manual/im-osqi-2019-01-review-and-approval-scientific>.)
2. Stay involved with the USGS software community to keep up to date with changing policies and procedures.
3. There are three components to the USGS git hosting platform, each with unique features and purposes: External, OpenSource, and InnerSource (<https://www.usgs.gov/products/software/software-management/develop-software>).

Tools You Can Use: Visualization, CDI Risk Map, TerriaMap, and More (Henry)

1. There are resources for displaying geospatial data and nongeospatial data (for example, TerriaMap, Tableau) that can be used in combination, and there are examples in the USGS (see the CDI Risk Map for one example); however, setting up these tools may still require web-developer or IT-administrator expertise.
2. Tools for building visualizations are great! BUT as we all start having the ability to build data visualizations, we need to consider the end users. Do they need a web mapper through which they can interact with everything, add data, and zoom infinitely? Or, does the end user need more of a guided experience to understand the volume of information we have to share? Additionally, consider whether a web page is necessary, or whether a still image or video file would be just as effective.
3. The next step is to continue the conversation about visualization tools available to CDI members at a future CDI monthly meeting.

Want to Know How to Make Your Data Applications Delightful? Usability Can Help! (Hou)

1. User-centered design and usability assessments will be critical for developing applications and dashboards that support “actionable intelligence” in the USGS.
2. Any usability testing is better than none.
3. Leverage different usability assessment techniques to test quickly and often.
4. Keep user tests focused and short.
5. Users sharing their thought processes aloud can be very helpful and useful as developers.
6. Multiple points of view (from potential users) can give good insight into the usability of your application (app).
7. Don’t waste the number of clicks from your users.

Appendix 4. Interactive Questions and Comments

At the workshop, audience feedback was collected and displayed in real time by the [sli.do](#) application (app). Participants used their mobile phones or laptops to submit questions or comments relevant to the current speaker or session. These questions and comments were then displayed on the user interface so that others using the app could upvote posts that they agreed with. During the plenary sessions, questions and comments with the most votes were raised for discussion. Other questions were selected for discussion during later CDI monthly meetings. In total, 141 users used the [sli.do](#) app, and 92 questions or comments were submitted. In addition, 245 votes were cast on nine polls administered through [sli.do](#).

Questions for CDI Executive Sponsors With the Highest Support

The opening and closing plenary sessions included segments that asked for questions to the CDI executive sponsors, Kevin T. Gallagher and Tim Quinn. The following topics had the highest support, measured by number of upvotes (shown in parentheses) on the [sli.do](#) app.

- How can we improve our USGS web presence in order to pursue the vision of USGS science priorities? (13)
- How does our workforce need to change in order to support the EarthMAP vision? (9)
- What is the vision for making unstructured or loosely structured data more structured and ready for integration in a data lake? (8)
- What are some incentives for producing or engaging in data integration and archiving? (7)
- How can we be receptive to new technology initiatives, such as edge computing and decision-support tools, consistently across the USGS? (7)

Selected Poll Responses From the Closing Plenary Session

Two questions were posed to the audience during the closing plenary session in order to guide future CDI programming and strategy: “What new information did you learn and will take back to your job?” and “What topic from the workshop requires further discussion within the CDI?” The workshop organizers received 57 and 54 responses to these questions, respectively. The following responses were selected to represent the variety of responses that are useful for CDI programming and strategy. The responses are listed in the order that they were received rather than in any ranked order.

What new information did you learn and will take back to your job?

1. I learned about tools (software, APIs) and methods that other data managers are using to meet the demands of their science centers—like the ScienceBase Python API for data releases.
2. I learned quite a bit about integrating models and using semantic ontologies to do so.
3. Our group can perform all of our modeling tasks on Denali, Tallgrass, and Black Pearl.
4. How to create a napkin sketch and value statement. How to think about the Who and Why before the What and How.
5. Connections to others dealing with similar issues and the need for a directory of software developers. This directory would help as we deal with similar development issues.
6. I learned how to leverage Python for data-management workflows. I also have a better understanding about how to release software to the public.
7. The importance of FAIR, usability studies, and user-centric design.
8. I’m now more aware of opportunities and ways to move into the cloud.

What topic from the workshop requires further discussion in the CDI?

1. How about a webinar that unpacked the linguistic hype around data “lake”? The webinar could discuss what lakes have in common with structured data repositories and authoritative data sources, who needs to understand the distinction, and when it is it important to differentiate, and when it is not.
2. Meet with quality assurance folks to develop data quality metric indicators to increase data confidence.
3. List training opportunities for software development and new scientific computing tools.
4. Unify our approach to data delivery in the forms our stakeholders need.
5. Software needs to be treated equally with data.
6. Diversity and inclusion.
7. Sharing tools, algorithms, and methods for predictive models.
8. Accessible training in data management for people newly in a data-management role. Integrating data, code, narrative, and web in the product review/publication process.
9. Use of biological data standards.
10. The Data Ops group idea is a good one and should be formed to help define data lake requirements.
11. We need to be very careful how we talk about visualizations. It’s easy to get excited about a tool, but users’ needs must come first, and the visualizations need to be made with good design and user-interaction principles in mind.
12. Metadata for newbies.
13. Curating big data from start to finish.
14. Usability—Bringing into the USGS more experts in UX/UI to help build capacity at science centers and to provide services directly for projects.
15. In-depth tech training (on data management, HPC, coding, and other skills) for new scientists/employees as well as continuing education for the rest of us.
16. Connecting developers with product owners more effectively to improve data product functionality.
17. Importance of software in data and science reproducibility.
18. Solutions for technical challenges related to collaborations with organizations outside of the USGS.
19. Data Science/Machine Learning/AI.
20. Data Visualization Community of Practice (or at least a webinar series). Focus on not just tools, but also best practices.
21. Transition to ISO metadata.
22. Coproduction.
23. Role of CDI in forecasting—Where can CDI aid in technological advances and integration of data and models?

Appendix 5. Community for Data Integration and Science Support Framework

[First published in Chang and others (2015). Minor edits were made to fit the format of this report.]

In order to provide an overarching context and vision for Community for Data Integration (CDI) goals and activities, the CDI coordinators, consisting of working group leads and facilitators, developed the Science Support Framework (SSF) in 2012. The SSF categorizes the activities and processes through which research data flow and upon which the CDI operates. It is these categories that provide the operational foundation and conceptual architecture that illustrate how CDI activities contribute to Bureau-level data integration efforts.

The vertical elements in the SSF (fig. 7) represent the “how” of the CDI—processes, implementation of standards and best practices, and interactions among people, data, and technology necessary to achieve data integration. The activities of monitoring, assessment, and research flow through the Science Data Lifecycle Model (Faundeen and others, 2013) process with the aid of applications, web services, and semantics (that is, common frameworks and ontologies for sharing data across applications, communities, enterprises, and so forth). The assets are transformed into information products that increase knowledge and understanding of the Earth’s physical and biological systems.

The horizontal elements in the SSF (fig. 7) represent the “what” of the CDI: products and tools and the mechanisms that mediate and contribute to the discovery and effective use of scientific data in systematic research. Data assets are managed within the context of the individual science projects, flowing horizontally from science project support through the Science Data Lifecycle Model processes, applications, and ultimately to data and knowledge management.

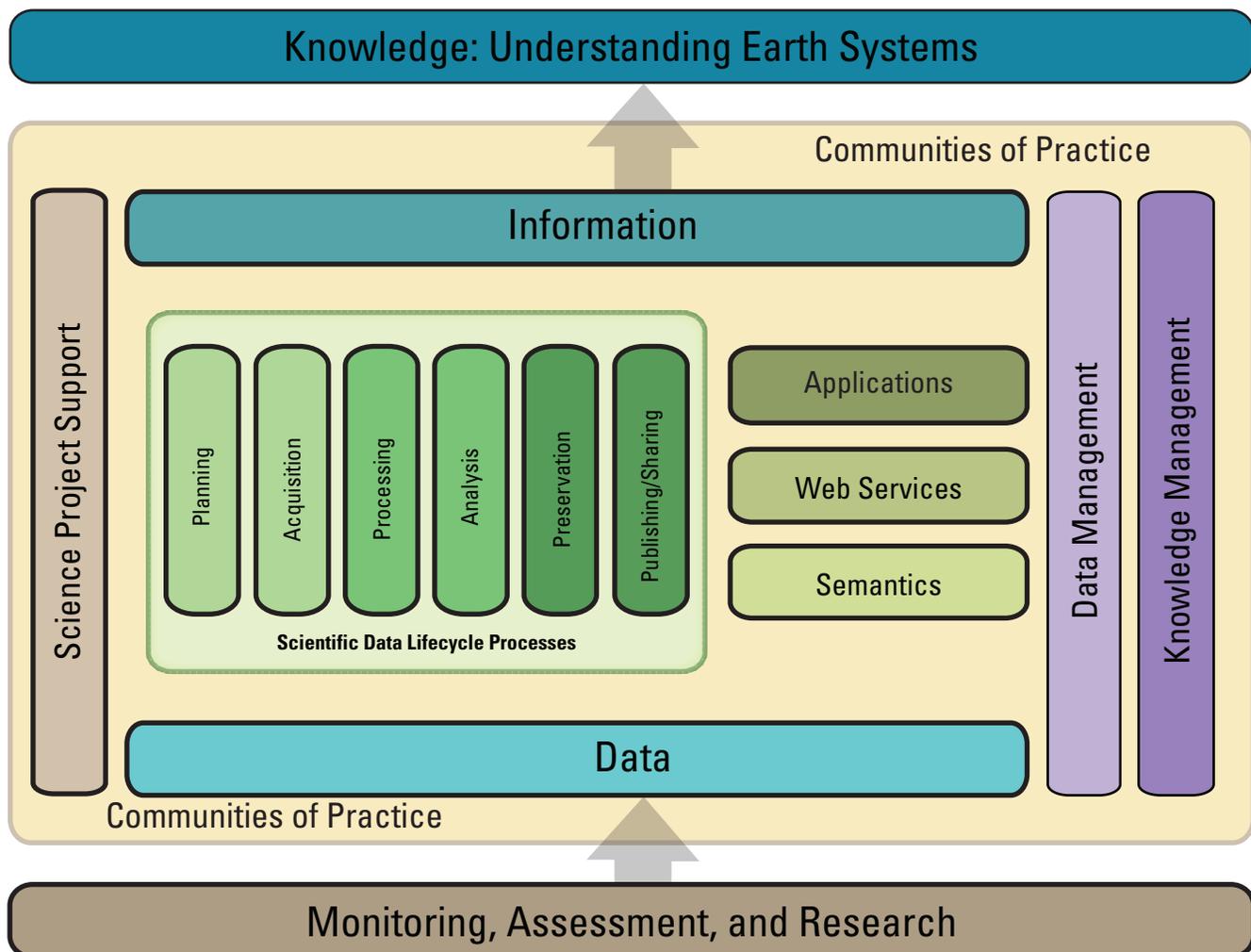


Figure 7. The Community for Data Integration Science Support Framework.

