USGS
science for a changing world

Community for Data Integration 2016 Annual Report

Open-File Report 2017–1053

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

Community for Data Integration 2016 Annual Report

By Madison L. Langseth, Leslie Hsu, Jon Amberg, Norman Bliss, Andrew R. Bock, Rachel T. Bolus, R. Sky Bristol, Katherine J. Chase, Theresa M. Crimmins, Paul S. Earle, Richard Erickson, A. Lance Everette, Jeff Falgout, John L. Faundeen, Michael Fienen, Rusty Griffin, Michelle R. Guy, Kevin D. Henry, Nancy J. Hoebelheinrich, Randall Hunt, Vivian B. Hutchison, Drew A. Ignizio, Dana M. Infante, Catherine Jarnevich, Jeanne M. Jones, Tim Kern, Scott Leibowitz, Francis L. Lightsom, R. Lee Marsh, S. Grace McCalla, Marcia McNiff, Jeffrey T. Morisette, John C. Nelson, Tamar Norkin, Todd M. Preston, Alyssa Rosemartin, Roy Sando, Jason T. Sherba, Richard P. Signell, Benjamin M. Sleeter, Eric T. Sundquist, Colin B. Talbert, Roland J. Viger, Jake F. Weltzin, Sharon Waltman, Marc Weber, Daniel J. Wieferich, Brad Williams, Lisamarie Windham-Myers

Open-File Report 2017–1053

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

U.S. Department of the Interior
RYAN K. ZINKE, Secretary

U.S. Geological Survey
William H. Werkheiser, Acting Director

U.S. Geological Survey, Reston, Virginia: 2017

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 http://www.usgs.gov or call 1–888–ASK–USGS.

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

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

Suggested citation:
Langseth, M.L., Hsu, Leslie, Amberg, Jon, Bliss, Norman, Bock, A.R., Bolus, R.T., Bristol, R.S., Chase, K.J., Crimmins, T.M., Earle, P.S., Erickson, Richard, Everette, A.L., Falgout, Jeff, Faundeen, J.L., Fienen, Michael, Griffin, Rusty, Guy, M.R., Henry, K.D., Hoebelheinrich, N.J., Hunt, Randall, Hutchison, V.B., Ignizio, D.A., Infante, D.M., Jarnevich, Catherine, Jones, J.M., Kern, Tim, Leibowitz, Scott, Lightsom, F.L., Marsh, R.L., McCalla, S.G., McNiff, Marcia, Morisette, J.T., Nelson, J.C., Norkin, Tamar, Preston, T.M., Rosemartin, Alyssa, Sando, Roy, Sherba, J.T., Signell, R.P., Sleeter, B.M., Sundquist, E.T., Talbert, C.B., Viger, R.J., Weltzin, J.F., Waltman, Sharon, Weber, Marc, Wieferich, D.J., Williams, Brad, Windham-Myers, Lisamarie, 2017, Community for Data Integration 2016 annual report: U.S. Geological Survey Open-File Report 2017–1053, 40 p., https://doi.org/10.3133/ofr20171053.

ISSN 2331-1258 (online)

Contents

Figures

Tables

Abbreviations

API	application programming interface
BAP	Bioscape Analysis Package
CCS	crowdsourcing and citizen science
CDI	Community for Data Integration
CDWG	Connected Devices Working Group
CWG	Communication Working Group
CZML	Cesium Markup Language
DCMI	Dublin Core Metadata Initiative
DEM	digital elevation model
DMP	data management plan
DMT	data management training
DMWG	Data Management Working Group
DOI	U.S. Department of the Interior
EMSC	European-Mediterranean Seismological Centre
EPA	Environmental Protection Agency

ESIP	Earth Science Information Partners
ETWG	Earth-Science Themes Working Group
EVT	ecosystem valuation toolkit
FAQ	frequently asked question
FGDC	Federal Geographic Data Committee
FORT	USGS Fort Collins Science Center
FSP	Fundamental Science Practices
FY	fiscal year
GAP	Gap Analysis Program
GIF	Geospatial Innovation Facility
HTTPS	HyperText Transfer Protocol Secure
iRIC	International River Interface Cooperative
IM	Instructional Memo
IT	information technology
JSON	JavaScript Object Notation
LRMI	Learning Resource Metadata Initiative
NASA	National Aeronautics and Space Administration
NBM	National Biogeographic Map
NEIC	National Earthquake Information Center
NHDPlus	National Hydrography Dataset Plus
NOAA	National Oceanic and Atmospheric Administration
NPN	National Phenology Network
NWI	National Wetlands Inventory
OFR	open-file report
OGC	Open Geospatial Consortium
OMB	Office of Management and Budget
OPeNDAP	Open-Source Project for a Network Data Access Protocol
ORNL DAAC	Oak Ridge National Laboratory Distributed Active Archive Center for Biogeochemical Dynamics
OSTP	Office of Science and Technology Policy
PI	principal investigator
RFP	request for proposals
SAHM	Software for Assisted Habitat Modeling
SOI	statement of interest
SSF	Science Support Framework
SSURGO	Soil Survey Geographic Database

STATSGO2	State Soil Geographic dataset
SWWG	Semantic Web Working Group
TED	Tweet Earthquake Dispatch
TSWG	Technology Stack Working Group
UC	University of California
UI	user interface
UMESC	USGS Upper Midwest Environmental Sciences Center
USGS	U.S. Geological Survey
WCS	Web Coverage Service
WFS	web feature service
WGS	World Geodetic System
WMS	web map service

Community for Data Integration 2016 Annual Report

By Madison L. Langseth, Leslie Hsu, Jon Amberg, Norman Bliss, Andrew R. Bock, Rachel T. Bolus, R. Sky Bristol, Katherine J. Chase, Theresa M. Crimmins, Paul S. Earle, Richard Erickson, A. Lance Everette, Jeff Falgout, John L. Faundeen, Michael Fienen, Rusty Griffin, Michelle R. Guy, Kevin D. Henry, Nancy J. Hoebelheinrich, Randall Hunt, Vivian B. Hutchison, Drew A. Ignizio, Dana M. Infante, Catherine Jarnevich, Jeanne M. Jones, Tim Kern, Scott Leibowitz, Francis L. Lightsom, R. Lee Marsh, S. Grace McCalla, Marcia McNiff, Jeffrey T. Morisette, John C. Nelson, Tamar Norkin, Todd M. Preston, Alyssa Rosemartin, Roy Sando, Jason T. Sherba, Richard P. Signell, Benjamin M. Sleeter, Eric T. Sundquist, Colin B. Talbert, Roland J. Viger, Jake F. Weltzin, Sharon Waltman, Marc Weber, Daniel J. Wieferich, Brad Williams, Lisamarie Windham-Myers

Abstract

The Community for Data Integration (CDI) represents a dynamic community of practice focused on advancing science data and information management and integration capabilities across the U.S. Geological Survey and the CDI community. This annual report describes the various presentations, activities, and outcomes of the CDI monthly forums, working groups, virtual training series, and other CDI-sponsored events in fiscal year 2016. The report also describes the objectives and accomplishments of the 13 CDI-funded projects in fiscal year 2016.

Introduction

The Community for Data Integration (CDI) represents a dynamic community of practice focused on advancing science data and information management and integration capabilities across the U.S. Geological Survey (USGS) and the CDI community. The CDI fosters an environment for collaboration and sharing by bringing together expertise from external partners and representatives across the USGS who are involved in research, data management, and information technology. Membership is voluntary and open to USGS employees and other individuals and organizations willing to contribute to the community (if interested, contact cdi@usgs.gov). In fiscal year (FY) 2016, the CDI welcomed 134 new members who are interested in learning from and sharing with the community.

The goals of the CDI are to

- advance understanding of Earth systems through enhanced use of data and information,
- provide a forum for data practitioners to share ideas and learn new skills and techniques, and
- grow USGS data and information capabilities by increasing the visibility of data integration work across the USGS and the CDI.

To achieve these goals, the CDI focuses on activities within five applied areas: monthly forums, annual workshop/webinar series, working groups, projects, and special events. The monthly forums provide an open dialogue to share and learn about data integration efforts or to present problems that invite the community to offer solutions, advice, and support. The CDI's annual workshop and webinar series bring community members together to share ideas and increase visibility of current projects and activities, as well as to provide training on state-of-the-art technologies and concepts. Stemming from common interests, the CDI working groups focus on efforts to address data management and technical challenges, including the development of standards and tools, improving interoperability and information infrastructure, and data preservation within USGS and its partners. Through the formal request for proposals (RFP) process, the CDI funds projects that produce tangible data-integration-related products to advance science and technology across the USGS and the Earth and biological science community. The CDI's Executive Sponsors, Kevin Gallagher (Associate Director, Core Science Systems) and Tim Quinn (Chief, Office of Enterprise Information), provide guidance, contribute funding, and advocate for the CDI's activities and projects. The purpose of this annual report is to inform the public about the outcomes of these activities and projects for FY 2016.

Monthly Forums

Every month, the CDI gathers for a virtual meeting forum. Monthly forums enable community members to stay up to date on new tools, best practices, standards, and policies within the Earth and biological sciences community. The CDI members and nonmembers alike are invited to give presentations on topics related to data integration. Table 1 lists the presentations from FY 2016. During these monthly forums, community members are encouraged to ask questions, present challenges, and share solutions to data integration problems. The monthly forums also provide the CDI Executive Sponsors and Coordinators with the opportunity to announce upcoming CDI activities and interact directly with the community. Additionally, the CDI working group leads are able to report progress on their activities during these meetings. In FY 2016, an average of 65 people attended each meeting (table 1).

Table 1. Monthly Community for Data Integration forum presentations for fiscal year 2016.

[CDI, Community for Data Integration; USGS, U.S. Geological Survey; FY, fiscal year; RFP, request for proposal; EPA, Environmental Protection Agency]

Date	Presentation title	Speaker(s)	Number of attendees
October 14, 2015	Public Lab	Mathew Lippincott, Public Laboratory for Open Technology and Science	38
	Citizen Sensing and the Problems and Practices of Citizen-Gathered Data	Jennifer Gabrys, Goldsmiths, University of London	
	Water Canary	Sonaar Luthra, Water Canary	
November 11, 2015	Canceled due to Veterans Day		
December 9, 2015	Engaging Citizens and Communicating Science through Open Innovation at the U.S. Geological Survey	Sophia Liu, USGS	50
January 13, 2016	CDI FY15 Project Presentation: CDI Land Cover Trends Photo Project	Chris Soulard, USGS	46
	CDI FY15 Project Presentation: CDI Geographic Searches Project	Rex Sanders, USGS	
February 10, 2016	CDI FY15 Project Presentation: The USGS dam removal information portal (DRIP)	Jeff Duda, USGS	63
	Visualizing USGS Science Data—the USGS Science Data Catalog and the Open Data Ecosystem	Ben Wheeler, USGS	
March 9, 2016	CyberGIS and CEGIS	Mike Finn, USGS, and Johnathan Rush, University of Illinois	87
	CDI FY15 Project Presentation: Digital Grain Size App	Daniel Buscombe, USGS	
April 13, 2016	No monthly meeting due to Software and Data Carpentry Workshops (see Special Workshops and Training Events section)		
May 11, 2016	CDI FY15 Project Presentation: sbtools: Connecting data to scientific computing	Luke Winslow, USGS	58
June 8, 2016	CDI FY15 Project Presentation: USGS Unmanned Aircraft System (UAS) Data Management Opportunities and Challenges	Brent Johnson, USGS	54
	Introduction to the CDI 2016 Virtual Trainings	Leslie Hsu, USGS	
July 13, 2016	USGS Cloud Hosting Solutions (CHS) Progress and Activities	Kimberly Scott, USGS, and Vickie Backus, USGS	98
	CDI FY15 Project Presentation: Implementing Controlled Vocabulary Services in USGS	Fran Lightsom, USGS; Peter Schweitzer, USGS; and Alan Allwardt, USGS	
August 10, 2016	Public Access / Open Access at USGS	Viv Hutchison, USGS	81
	Data-Driven Discovery	Carly Strasser, Data-Driven Discovery Initiative	
September 14, 2016	CDI FY17 RFP Announcement	Kevin Gallagher, USGS, and Tim Quinn, USGS	70
	Citizenscience.gov Resources and Paperwork Reduction Act Update	Alison Parker, EPA, and James Sayer, USGS	
	CDI in FY17—Interactive polling and feedback for the Annual Meeting and More	Leslie Hsu, USGS	

Since March 2016, the CDI monthly forums have included a new segment called "Scientist's Challenge." The Scientist's Challenge is a short block of time at the beginning of each meeting for scientists to present challenging problems that they are working to solve. The purpose is to tap into the CDI's powerful collective body of knowledge to form connections and possible future collaborations across the USGS and the Earth and biological sciences community. The following are some of the challenges that have been presented.

- What are approaches for developing a web application programming interface (API) to access, subset, and project digital elevation models (DEMs) and imagery available at the USGS Open-Source Project for a Network Data Access Protocol (OPeNDAP) and serve the data directly into the International River Interface Cooperative (iRIC) application?
- What tools and methods are available for archiving model software and data?
- In cross-disciplinary hazards research, seismic wave data are applied to processes besides earthquakes, like landslides, debris flows, and floods. What is the best way to catalog the diverse data types (seismic, geographic information system [GIS], imagery, text) for efficient discovery and analysis?
- What is the best way for a diverse group of geochronological and geochemical labs to embark on improved and efficient data management that will aid their data releases?

Each Scientist's Challenge is posted to the CDI forum, and community members are able to reach out to the scientists and provide guidance, resources, and collaboration opportunities. Outcomes and solutions are also documented on the CDI forum. The most common solution has been to initiate a more detailed discussion with a CDI member who has previous experience with the topic. Challenges have also developed into more mature CDI statements of interest for the annual RFPs.

Virtual Training Series

Following the 2015 CDI Workshop (Langseth and others, 2016), the community decided to schedule in-person workshops every other year while holding a virtual event in the off years. This year, the CDI held the 2016 Virtual Training Series. Initial ideas for training and presentations were generated from the CDI IdeaLab (a forum for posting and voting on ideas), workshop surveys, and CDI Coordinator discussions. The final selection of topics was chosen by an informal community voting process. The 2016 Virtual Training Series featured the following three training events: Reviewing Metadata and Using Controlled Vocabularies, Git and Bitbucket—Version Control, and Scientific Workflow and Reproducibility. Each training session is described in more detail in the following sections, and training materials and discussion are available on the CDI wiki at https://my.usgs.gov/confluence/x/ZsBtI.

Reviewing Metadata and Using Controlled Vocabularies

The Reviewing Metadata and Using Controlled Vocabularies session was divided into two separate events. The first event on using controlled vocabularies was presented at the July monthly forum and included demonstrations of USGS vocabulary services and metadata tools that use the vocabulary services. Developing these vocabulary services increases the ease and use of controlled vocabularies, which help to improve the discovery, access, use, and integration of data. Ninety-eight people participated in this monthly forum.

The second event on reviewing metadata took place in August 2016. During the training session, the trainers demonstrated how to perform primary validation on the Federal Geographic Data Committee's (FGDC) Content Standard for Digital Geospatial Metadata records using the Metadata Parser (Schweitzer, 1995). Ensuring that metadata robustly describe data is essential for future reuse and integration with other datasets. During this event, the trainers discussed techniques for reviewing the metadata for robustness after validation and concluded with approaches for examining the data in conjunction with the metadata. Seventy-five people participated in the reviewing metadata section of this training event. A major outcome of this training event was the development of the CDI Metadata Reviewers Community of Practice, which started meeting monthly in September 2016 to discuss and learn about best practices, resources, and Bureau-wide challenges regarding metadata review.

Git and Bitbucket—Version Control

The FY 2016 CDI projects were required to ensure source code developed during the course of the project was preserved in the version control system, the USGS Bitbucket Repository. This requirement was established by the CDI Coordinators in anticipation of the USGS software release policy that was being drafted in FY 2016. To help CDI project principal investigators (PIs), as well as others, learn how to comply with these new requirements, CDI members came together to develop a training session on version control applications such as Git and Bitbucket. The session consisted of an introduction to the benefits of using Git and version control and a demonstration of how to use Git BASH, SourceTree (a Git graphical user interface), and Bitbucket to clone, commit, fetch, pull, and push source code to and from a repository. The training session wrapped up with a brief overview of the draft USGS software release policy. Sixty people participated in this virtual training event.

Scientific Workflow and Reproducibility

The final event in the FY 2016 Virtual Training Series was Scientific Workflow and Reproducibility. This session was led by April Clyburne-Sherin from the Center for Open Science. The objectives of this session were for participants to understand the current issues and barriers to reproducibility in science, how the complete scientific workflow can affect reproducibility, and what documentation and organizational issues underpin reproducibility. The training and demonstrations featured the Center for Open Science's Open Science Framework tool. Fifty people participated in this virtual training event.

Special Workshops and Training Events

In addition to the Virtual Training Series that took place in July and August 2016, four other workshops and conference sessions were also coordinated by the CDI. The four events were Software and Data Carpentry workshops, a Data Hackathon, a virtual Mapping Innovation Workshop, and a SciDataCon 2016 session. These events were all coordinated at the request of CDI members and often in partnership with other groups. The sections below describe the events in more detail.

Software and Data Carpentry

The CDI partnered with the Federation of Earth Science Information Partners (ESIP) to jointly host in-person Software and Data Carpentry workshops at the Denver Federal Center. The events were held to meet the demand for scientific programming and data handling skills using open-source languages. The Software and Data Carpentry workshops were held concurrently over a 2-day time period and were taught by experienced Software and Data Carpentry instructors. Volunteers for the workshops were recruited through the CDI and ESIP email lists.

Software Carpentry

The Software Carpentry workshop was geared to help two groups of scientists:

- scientists who did not know how to program but wanted to learn to make their research more efficient and reproducible, and
- scientists who were programming but doing so inefficiently.

The Software Carpentry workshop was taught using Python; however, the instructors emphasized that the language was not as important as the programming skills. The goal of the workshop was to teach participants how to organize their data and workflow and modularize their code to make them more efficient at programming. The lessons learned in this workshop were transferrable to any programming language. The workshop was divided into four categories: automating tasks with the Unix shell, handling data with Python, building programs with Python, and version control with Git.

Data Carpentry

The Data Carpentry workshop was geared to teach people who were managing data manually about how to automate their work and make it more reproducible. The Data Carpentry workshop incorporated new Data Carpentry lessons that aimed to teach scientists how to perform geospatial analysis in the R programming language. The instructors started off with an introduction to geospatial concepts such as spatial data formats, metadata formats, and structure and then moved into working with vector data in R. Participants learned how to open and plot shapefiles, explore shapefile attributes, handle spatial projection and coordinate reference systems, and convert from a comma-separated values file to a shapefile in R. Participants also learned how to work with raster data, including plotting, reprojecting, and cropping raster data.

2016 Data Hackathon

Following the 2-day Software and Data Carpentry workshops, CDI and ESIP hosted a data hackathon with the goal of providing CDI and ESIP members with an opportunity to creatively work with USGS data in an environment conducive for networking among USGS employees and external participants. The hackathon provided Software and Data Carpentry participants with an opportunity to use their newly acquired skills on a set of issues surrounding USGS datasets and tools. The hackathon topics included improving accessibility of datasets through APIs (Ocean Biogeographic Information System USA dataset); scraping, compiling, and standardizing data files listed on websites (Critical Minerals datasets); and accessing the USGS ScienceBase Repository using the ScienceBase Python library, pysb. Participants were given a choice to work on one of these topics or other personal data challenges. Code developed and used during the hackathon was made accessible via tools like iPython notebooks, GitHub, and Bitbucket and documented on the Hackathon HackPad (https://usgs-esip-hackathon.hackpad.com/).

Hackathon participants were enthusiastic about the event and when asked, “How can we continue the momentum of this first hackathon?” participants replied, “Hold more hackathons! Regularly!” The CDI hopes to host more data hackathons in the future and capitalize on this exciting, collaborative, real-time method of finding solutions to shared challenges (figure 1).

Figure 1. Photographs showing Data Hackathon participants working to solve data challenges.

Mapping Innovation Workshop

In August and September 2016, the USGS hosted a number of in-person and virtual Mapping Innovation Workshops. The goal of the workshops was to start a conversation about the opportunities and challenges of mapping in the USGS. The CDI hosted one of the five virtual workshops with approximately 50 participants from all USGS mission areas and regions except for Alaska.

Prior to the workshop, mapping innovation stories were solicited from the participants, with a focus on CDI-funded projects. Most registrants marked that they wanted to join the Mapping Innovation Workshop to “learn about innovative mapping tools and resources,” and the presentation of these stories helped to fulfill that need. A wide range of challenges was mentioned in the mapping innovation stories, from dealing with disparate data types to finding available programmers and understanding USGS Fundamental Science Practice requirements.

The discussion portion of the workshop revolved around three main themes: data interoperability, data discovery and sharing, and communication and visualization. Google forms were used to gather responses from the virtual participants. The facilitators discussed the submitted answers and asked for clarification when necessary. One of the major challenges that surfaced at this workshop was the financial and educational challenge of making data as interoperable as possible. Participants also identified topics for further discussion at future CDI monthly or annual meetings, including use of the semantic web to make data discoverable and new mapping tools such as Leaflet and D3.

SciDataCon 2016

CDI members coordinated the USGS participation in SciDataCon, a conference seeking to advance the frontiers of data in all areas of research, held in Denver in September 2016. Members of the CDI coordination team organized USGS participant interest in the SciDataCon theme of “Policy and Practice of Data in Research.” CDI members Madison Langseth and John Faundeen convened the session “Managing Science Data: A Federal Agency’s Perspective.”

The session had seven accepted oral presentations and one poster. Six of the seven presentations were given by USGS representatives on the topics of the USGS Data Management website, the CDI community, monitoring tools, trusted digital repositories, metadata management, and data management strategies (table 2). The session had over 50 attendees and helped to spread the word about USGS data management successes to the international scientific data community. Papers from the proceedings are archived at https://my.usgs.gov/confluence/x/naReIQ.

Table 2. SciDataCon 2016 session “Managing Science Data: A Federal Agency’s Perspective.”

[USGS, U.S. Geological Survey; NIST, National Institute of Standards and Technology]

Presentation name	Presenter
USGS Data Management Website: Helping Our Scientists	Michelle Chang, USGS
MonitoringResources.org: A Suite of Online Tools to Document Monitoring Protocols, Methods, and Designs to Promote Data Sharing and Preservation	Rebecca Scully and Jennifer Bayer, USGS
USGS Community for Data Integration (CDI): A Community Approach to Managing Science Data	Leslie Hsu, USGS
Developing Criteria to Establish Trusted Digital Repositories	John Faundeen, USGS
Metadata Management Implementation in a Large Federal Organization	Raymond Obuch, USGS
Getting Started With Data Governance in a Wide Agricultural Research Organization: Challenges, Opportunities and Ways Forward	Debora Pignatari Drucker, Embrapa
Data Management Challenges in a Distributed Organization: What Challenges We Are Facing at the USGS and How Are We Working to Overcome Them	J.C. Nelson, USGS
Research Data Management at the National Institute of Standards and Technology: Making Data Openly Accessible (Poster)	Regina Avila, NIST

Working Groups and Focus Groups

The CDI is organized into working groups that form around common interests in specific topics related to data integration (table 3). These working groups provide a platform for sharing resources and knowledge, discussing challenges, and identifying solutions that will help to advance data integration in the Earth and biological sciences. Some working groups meet on a regular basis, whereas others meet when the need arises. Each working group has one or more leaders to coordinate meetings, projects, and information sharing as well as to report current activities up to the larger CDI community. Working group membership is voluntary and open to anyone interested in participating.

Table 3. Community for Data Integration working groups and contacts.

Working group name	Working group contact(s)
Citizen Science Working Group	Sophia Liu – sophialiu@usgs.gov
Communication Working Group	John C. Nelson – jcnelson@usgs.gov Marcia McNiff – mmcniff@usgs.gov
Connected Devices Working Group	Tim Kern – kernt@usgs.gov Lance Everette – everettel@usgs.gov
Data Management Working Group	Viv Hutchison – vhutchison@usgs.gov
Earth-Science Themes Working Group	Roland Viger – rviger@usgs.gov
Semantic Web Working Group	Fran Lightsom – flightsom@usgs.gov
Technology Stack Working Group	Richard Signell – rsignell@usgs.gov

Citizen Science Working Group

The Citizen Science Working Group communicates relevant information about citizen science and crowdsourcing to interested parties. Crowdsourcing and citizen science (CCS) are important methods that the USGS can use to engage and inform the public and even gain valuable data observations from them. The activities described here significantly advance the USGS mission to monitor, assess, and conduct targeted research. As outlined in an August 2016 USGS Leader's Blog post, Sophia Liu, the coordinator for this working group, is working to build a Steering Committee for Open Innovation that focuses on crowdsourcing, citizen science, and civic hacking (https://my.usgs.gov/confluence/x/GgD1I).

In FY16, there were two CDI Monthly Meetings focused on citizen science that summarized CCS opportunities in the USGS and Federal Government and the efforts to make citizen science projects easier to initiate (December 2016 and September 2016). The website https://www.citizenscience.gov was launched in April 2016, and the CDI working group supported communication with USGS PIs during the compilation and vetting of project information for the site. The catalog at https://www.citizenscience.gov is a definitive inventory of Federal crowdsourcing and citizen science projects, and the USGS is listed as an agency sponsor for 42 projects.

Communication Working Group

The Communication Working Group (CWG) started in FY 2016. The goal of the CWG is to create lines of communication between the CDI, the science centers, regions, and mission areas of the USGS. A large portion of the year was spent fine tuning the goals and purpose of the group. The hope for membership is to represent all of the USGS mission areas and regions. Having this diverse membership will allow the group to establish a more robust and effective communication network for the CDI.

Communication Working Group Accomplishments

This new working group was formed within the CDI to address communication needs identified at the May 2015 CDI face-to-face meeting. The working group leads, along with a small group of others, worked with Cheryl Morris, Director of the USGS Core Science Analytics, Synthesis and Library, to develop a strategy to solicit executive buy-in to phase out the Science Data Coordinator Network and replace it with the CWG, which would serve a similar purpose but would reside within the CDI. Initial activities included focusing and prioritizing the CDI Communication Plan for FY 2017 and evaluating a communication activity matrix for CDI.

Membership includes people from the USGS Midwest, Southwest, and Northwest Regions and from the Ecosystems and Core Science Systems mission areas. In the year ahead, the group hopes to expand its membership to include representation from all USGS regions and mission areas so as to better disseminate CDI information.

Connected Devices Working Group

The Connected Devices Working Group (CDWG), as a community of practice, focuses on being both a learning environment and a forum for members to exchange ideas. The group supports the following:

- information technology (IT) staff working with vendors to gain insight into new technologies being developed,
- software developers by promoting coding and design best practices and by discussing technical and design issues, and
- scientists and researchers who want to find out about technologies they can employ in their research.

The CDWG tries to facilitate collaboration among these different groups in an effort to promote joint projects, suggest technologies that can help with research goals, identify cost efficiencies and technical options, and introduce new concepts and discoveries with mobile technologies.

Connected Devices Working Group Accomplishments

In FY 2016, the CDWG had a number of accomplishments. The group established and managed the USGS iOS (Apple) App Store space and set up a Test Flight, which is a way for USGS projects to test iPhone applications. The CDWG established a review process for mobile applications, which is tied to the Mobile Application Release Checklist (Kern, 2015). They also reviewed and commented on the Department of Interior's Mobile Privacy Policy (Burns, 2016) and communicated information from the DigitalGov Mobile and Internet of Things calls.

Connected Devices Working Group Meetings and Presentations

There were five CDWG meetings in FY 2016, each including presentations designed to facilitate a discussion of specific software development issues (table 4). Complete meeting agendas and presentations are available at the CDWG Confluence site: https://my.usgs.gov/confluence/x/L4zyHw.

Table 4. Connected Devices Working Group meetings and presentations for fiscal year 2016.

[USGS, U.S. Geological Survey]

Date	Title	Speaker
January 2016	USGS Mobile Release	Tim Kern, USGS
February 2016	ScienceCache	Megan Eberhardt-Frank and Thom Miller, USGS
May 2016	Draft USGS IM on Software Release Environmental Protection Agency Generic Clearance for Citizen Science Data Collection	Tim Kern, USGS David Govoni, USGS
June 2016	Using Citizen Science to Measure Fisheries Harvest in Developing Countries Overview of Angular2, Ionic2, and Electron2 for Building Web, Mobile, and Desktop Apps from One Set of Typescript Components	Abigail Lynch and Bonnie Myers, USGS Thom Miller, USGS
August 2016	National Map Corp Mobile Effort Software Development USGS and Department of Interior Policy Updates	Elizabeth McCarthy and Rachel Stevenson, USGS Dell Long, USGS Tim Kern, USGS

Data Management Working Group

Good data management is a prerequisite for data integration, and the Data Management Working Group (DMWG) seeks to develop mechanisms for incorporating data management into USGS science and advancing education about its long-lasting value. The group seeks to elevate the practice of data management such that it is seen as a critical part of the pursuit of science in the USGS. In FY 2016, the DMWG concentrated on supporting the implementation of the USGS Public Access Plan, addressing requirements for a data release to accompany the publication of a scholarly conclusion (U.S. Geological Survey, 2016).

Data Management Working Group Accomplishments

In support of the USGS Public Access Plan, the DMWG sponsored several working teams to accomplish its goals. A description of each team and its work is provided below.

Data Policy Team

The Data Policy Team reviewed and updated the four USGS data management-related Instructional Memos (IMs) in order to convert each to policy. The IMs were first published in February 2015 and were published as policy in FY 2017. The policies are as follows: Scientific Data Management Foundation; Metadata for Scientific Data, Software, and Other Information Products; Review and Approval of Scientific Data for Release; and Preservation Requirements for Digital Scientific Data.

USGS Public Access Plan Implementation Coordination

In FY 2016 the DMWG turned much of its attention to the requirements outlined in the USGS Public Access Plan. To this end, 30 people, representing many disparate areas of the USGS that support applications and processes related to data release and scholarly publication, came together in March 2016 for a 3-day meeting in Reston, Virginia. Outcomes of the meeting included a roadmap of interdependent tasks with timelines for completion, three working groups assigned to forwarding communication, connecting systems and applications, and defining criteria for trusted digital repositories in the Bureau. With a deadline of October 1, 2016, for implementation of the USGS Public Access Plan, the DMWG coordinated monthly accountability meetings open to anyone with an interest in progress and weekly update meetings intended for Plan coordinators.

Data Management Website

The DMWG initiated the USGS Data Management website (https://www.usgs.gov/datamanagement) in 2012 and continues to provide support to enhance the site. In 2016, the website was enhanced with the addition of content for two major elements of the science data lifecycle, on which the structure of the website rests. The “Process” and “Analyze” sections were added through the effort of a team of people knowledgeable about these areas in the science data lifecycle. Additionally, the DMWG added new content to the following sections of the website.

- Training and Resources: modules entitled “ScienceBase as a Platform for Data Release,” “USGS Science Data Lifecycle,” “Metadata for Research Data,” and “Planning for Data Management.”
- Publish/Share: Updated information on data release in USGS.
- Data Management Plans: Updated the data management checklist.
- Data Management Plans: Added the template for developing a science center data management plan.
- Data Management Plans: Added a Data Management Plan (DMP) Tool Comparison Chart
- Data Management Plans: Top Nine Best Practices and Frequently Asked Questions (FAQ).

Finally, the USGS Data Management website entered into a Memorandum of Agreement with the U.S. Air Force Research Laboratory to host instances of the website for their organization. As appropriate, content updates and information will be shared between the organizations.

Science Center Strategy Development Working Group

This group collaborated to develop a science center data management strategy template for Science Center Directors and staff to aid in covering each of the many considerations associated with support for the USGS Public Access Plan and data management in their center. The plan focuses on "providing a core adoptable approach to managing new data requirements and reducing burden on staff and operational redundancy." Additionally, the team developed an evaluation of current DMP tools and produced a top nine DMP best practices document to aid researchers in developing DMPs for research projects. The best practices document and the center strategy template are both posted on the USGS Data Management website.

Data Management Working Group Meetings and Presentations

In FY 2015, the DMWG partnered with the Pacific Northwest Monitoring Partnership to sponsor a Data Management Webinar Series. This series continued through February 2016. Additionally, the DMWG held monthly meetings that featured short presentations as detailed in table 5.

Table 5. Data Management Working Group webinar series and monthly meeting presentations for fiscal year 2016.

[USGS, U.S. Geological Survey]

Date	Title	Speaker
November 2015	Best Practices for Preparing Data to Share and Preserve (webinar series presentation)	Robert Cook, Oak Ridge National Laboratory
January 2016	Data Citation and You: Where Things Stand Today (webinar series presentation)	Ruth Duerr, National Snow and Ice Data Center
February 2016	Open Data and the USGS Science Data Catalog (webinar series presentation)	Ben Wheeler, USGS
February 2016	Using Microsoft Access for Data Processing	Steve Tessler, USGS
March 2016	Update on the USGS Public Access Plan Implementation Meeting Southeast Region Data Managers Group—Purpose, Products, and How to Get Involved	Viv Hutchison, USGS Cassandra Ladino, USGS
April 2016	Using ScienceBase for Data Release in USGS Science Center Strategy Development	Madison Langseth, USGS Cassandra Ladino, USGS
June 2016	Alaska Science Center Data Policies, Data Management Planning, Data Release Workflow, and Tracking System	Dennis Walworth, USGS
July 2016	A Quick Tour of the Data Release Workbench	Viv Hutchison, USGS
September 2016	New Process and Analyze Page on the Data Management Website	Steve Tessler, USGS, and Michelle Chang, USGS

Earth-Science Themes Working Group

The Earth-Science Themes Working Group (ETWG) is intended to provide a focal point for applied Earth science within the CDI. In addition to building social networks around community-defined research questions, the CDI hopes to communicate and inject new ideas and best practices, such as the Science Data Life Cycle (Faundeen and others, 2013) and modern scientific computing approaches, from other working groups into the more traditional project-based work of the USGS and other participants. An additional goal of the ETWG is to bring fundamental Earth science data producers, such as the USGS National Hydrography (NHD), 3D Elevation , and Multi-Resolution Land Characteristics Programs, into more direct and regular contact with scientists who work to integrate these sometimes independent data sources. While the ETWG provides an umbrella for distinct themes such as water, elevation, soils, land cover, and oceans, its real promise is in the Integration Focus Group. The Integration Focus Group, by providing a home for applied-science topics that do not fit neatly under individual data theme headings, brings data, methods, and expertise from multiple domains together. The synergy of this process has exciting potential for changing how scientists pursue Earth science.

Part of the CDI's success has been its ability to concentrate the thinking and voices of data scientists, technologists, and communication specialists in our community to the point that a consensus of thinking on issues important to the execution of Earth sciences has emerged. The ETWG represents an evolutionary step for the CDI by complementing CDI's current demographic with more traditional Earth science researchers who serve as a real-world user base that will not only "test" other CDI working group ideas, but also help identify Earth science data integration needs and topics. Although focus groups within ETWG are entirely ad hoc and informally pursued, it is hoped that with adequate coordination, these efforts can be groomed and channeled into connections with larger communities such as the Critical Zone Observatory, EarthCube, ESIP, or the USGS Powell Center. In FY 2016, themes within the umbrella group continued to build content and community. Specific examples of activity in the last year include water use and riparian mapping.

Semantic Web Working Group

The Semantic Web Working Group (SWWG) is a small group of data managers who are working together to explore semantic web technologies for use in their jobs and also to improve the discovery, access, use, and integration of USGS data. In the past, the SWWG hosted presentations and seminar-style discussions and sponsored two CDI-funded projects. FY 2016 was a more informal year, characterized by meetings at which the group experimented with semantic web technologies and shared news and advice about progress they were making, separately, toward shared goals.

Semantic Web Working Group Accomplishments

At the SWWG face-to-face meeting at the 2015 CDI Workshop, the group embarked on a practical learning project. The goal was to investigate the possibility of using a Geographic SPARQL Protocol and RDF Query Language (GeoSPARQL) endpoint to release foundational USGS data holdings as geospatially enabled linked data that can be used as a semantic framework for interdisciplinary data integration. During multiple monthly sessions, the SWWG experimented with a “sandbox” installation of the Parliament Triple Store. Finally, the group concluded that the computational overhead of GeoSPARQL is not a good investment unless the geospatial element of the data varies with time, which might be expected with biological or coastal data but not with geological or geographical data. For static geospatial data, it would be more efficient to use semantic reasoning processes with well-defined spatial relationships (such as “upstream of” or “contained within”) served by a basic SPARQL endpoint.

The working group produced an implementation plan at the conclusion of its 2014–15 CDI-funded project, “Use of Controlled Vocabularies in USGS Information Applications” (Lightsom and others, 2015). As the opportunity arises for working group members, progress is being made toward improving USGS metadata and the discovery of data in the USGS Science Data Catalog. Discussions during working group meetings have served to assist and coordinate these actions.

Technology Stack Working Group

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

Technology Stack Working Group Accomplishments

After a survey of working group participants, the TSWG agreed that webinars about emerging technologies are an effective and useful exercise for the group. After reviewing recent talks in the series, the group realized that the subjects being explored were not specific to the USGS and could be of benefit to a larger audience. After discussions with the ESIP, the TSWG decided that it could merge the existing ESIP “Rant and Rave” webinar series to form a new, combined webinar series. The two groups decided to call the new webinar series the ESIP “Tech Dive” series, and the TSWG lead, Rich Signell, became co-chair with Ethan Davis of Unidata and the ESIP Information Technology and Interoperability Committee. The ESIP Information Technology and Interoperability Committee web page is http://wiki.esipfed.org/index.php/Interoperability_and_Technology.

This merger has been extremely successful, increasing attendance from 3–5 participants to 20–30 participants, with a high of 65 attendees. Descriptions of the monthly meetings and presentations are provided in table 6. In addition, ESIP Tech Dive webinars have integrated audio, and recordings are immediately available after the meeting on YouTube.

Table 6. Technology Stack Working Group meetings and presentations for fiscal year 2016.

[NOAA, National Oceanic and Atmospheric Administration; NASA, National Aeronautics and Space Administration]

Date	Title	Speaker
October 22, 2015	Wakari Enterprise & Jupyterhub	Ian Stokes-Rees, Continuum Analytics
November 19, 2015	OpenClimate GIS	Ben Koziol, NOAA
January 21, 2015	Geonode, pycsw, and CKAN	John Jediny, Data.gov
February 18, 2016	Bird-House: Web Processing Services Made Easy	Carsten Ehbrecht, Deutsches Klimarechenzentrum
March 17, 2016	New Python Mapping Tools	Filipe Fernandes, South East Coastal Ocean Observing Regional Association
April 21, 2016	The New Geoplatform.gov	Tod Dabolt, Department of the Interior
May 12, 2016	Leaflet Time Dimension	Biel Frontera, Balearic Islands Coastal Observing and Forecasting System
June 9, 2016	Dive into Docker	Kyle Wilcox, Dave Foster, and Shane St. Clair, Axiom Data Science
July 13, 2016	The NOAA OneStop Data Discovery and Access Framework Project	Ken Casey, NOAA National Centers for Environmental Information
August 11, 2016	Community Data Analysis Tools (CDAT)	Charles Doutriaux, Lawrence Livermore National Laboratory
September 8, 2016	Apache Open Climate Workbench	Lewis McGibbney and Kyo Lee, NASA Jet Propulsion Laboratory

Annual Community for Data Integration Request for Proposals

The CDI seeks to build and share knowledge about topics such as data integration, data handling and stewardship, scientific computing, and approaches for knowledge delivery. The main goal of CDI funding is to improve our collective knowledge about how to create better, longer lasting, and more accessible science products by leveraging the tools, methods, and datasets available to the Earth and biological science communities. The CDI places high value on innovative projects that, in the near future, produce new and reusable ideas, methods, or tools that have an impact beyond a single USGS program, center, region, or mission area. CDI project proposals are evaluated based on their alignment with the CDI Science Support Framework (SSF) (U.S. Geological Survey, 2015), the evaluation criteria laid out in the RFP guidance document (scope, technical approach, project experience and collaboration, sustainability, budget justification, and timeline), and the following guiding principles:

- focus on targeted efforts that yield near-term benefits to Earth and biological science;
- leverage existing capabilities and data;
- implement and demonstrate innovative solutions, such as methodologies, tools, or integration concepts, that could be used or replicated by others at scales from project to enterprise;
- preserve, expose, and improve access to Earth and biological science data, models, and other outputs; and
- develop, organize, and share knowledge and best practices in data integration.

In 2014, the CDI established a two-phased RFP process. This two-phased approach provides more transparency and community participation in the selection process by inviting community members to vote on two-page statements of interest (SOIs) submitted by project PIs. The SOIs receiving the most votes from the community, as well as SOIs identified by the Executive Sponsor as addressing an emerging priority, are asked to submit a full proposal. Formal guidance for the FY 2016 RFP was released on September 9, 2015. The guidance document outlined a two-phased approach that would be used for selecting the CDI FY 2016 projects.

Phase I—Statements of Interest

Two-page SOIs were due on October 9, 2015. A total of 33 SOIs were submitted representing 14 SSF elements (table 7). The lead PIs on the SOIs represented six USGS mission areas (table 8) and all seven USGS regions (table 9).

The CDI community members were asked to read all 33 SOIs and vote on them based on the CDI SSF, the evaluation criteria, and the guiding principles previously described. The voting period began on October 19, 2015, and closed on November 3, 2015. Each community member was allowed 15 votes to use across all SOIs. Each SOI could receive a maximum of three

Table 7. Number of statements of interest addressing each Science Support Framework element for fiscal year 2016.

[Some proposals addressed more than one Science Support Framework element]

Science Support Framework element	Number of proposals
Applications	19
Data	13
Science data lifecycle	12
Data management	10
Information	9
Science project support	7
Web services	6
Processing	6
Analysis	6
Publishing/sharing	5
Semantics	2
Communities of practice	2
Preservation	1
Knowledge management	1

Table 8. Number of statements of interest by U.S. Geological Survey mission area for fiscal year 2016.

Mission area	Number of proposals
Water	12
Ecosystems	11
Climate and Land Use	5
Natural Hazards	3
Energy and Minerals	1
Core Science Systems	1

Table 9. Number of statements of interest by U.S. Geological Survey region for fiscal year 2016.

Region	Number of proposals
Midwest	9
Northwest	6
National	4
Northeast	3
Pacific	3
Southwest	3
Southeast	3
Alaska	1

votes per person. A closing session was held on November 3 to allow the community to agree on the number of SOIs that would be recommended to move forward to the full proposal phase of the RFP. During the closing session, the community agreed that at least the top 14 proposals should move on to the full proposal phase. Following the closing session, the CDI Coordinators also reviewed the SOIs and recommended that an additional five proposals move on to the next phase. In the end, 19 SOIs were approved by the Executive Sponsors to be invited to submit full proposals.

Phase II—Full Proposals

Full proposals were due on January 22, 2016. Two authors chose not to submit a full proposal; therefore, 17 full proposals were submitted for the second phase of the RFP process. The CDI convened a formal, 7-person review panel to evaluate the 17 full proposals. The reviewers were all USGS employees and volunteered their time to the review panel. The reviewers represented a wide range of USGS mission areas, regions, and programs and brought with them a variety of scientific and technical expertise. The review panel consisted of CDI and non-CDI members. The reviewers were responsible for disclosing any potential conflicts of interest and recusing themselves from discussions involving proposals in question. The reviewers were also asked not to divulge the identity of the other reviewers.

Each reviewer was assigned to lead the discussion of two or three proposals. The discussion leader was responsible for having in-depth knowledge of their assigned proposals; however, the reviewers were responsible for reading all 17 full proposals and providing a cursory evaluation of each proposal’s strengths and weaknesses. Reviewers scored each proposal based on the following weighted evaluation criteria.

- Scope (25 percent)
- Technical approach (25 percent)
- Project experience and collaboration (25 percent)
- Sustainability (15 percent)

- Budget justification (5 percent)
- Timeline (5 percent)

Each proposal was discussed in turn over the course of three review sessions, and reviewers were allowed to modify their scores based on the feedback of the other reviewers. During the final review session, the panel collectively discussed the proposal rankings and agreed upon a final recommendation for each proposal. The review panel agreed on an order of priority for the full proposals to be funded based on availability of funds.

Recommendations

The prioritized list from the CDI review panel was presented to the CDI Executive Sponsors, Kevin Gallagher and Tim Quinn, for final selection and approval. On March 9, 2016, Kevin and Tim announced funding 13 new projects (table 10). The "Community for Data Integration Projects" section describes the projects and their accomplishments in more detail.

Table 10. Overview of the Community for Data Integration request for proposals projects funded in fiscal year 2016 (in alphabetical order). Project title hyperlinks resolve to a ScienceBase item describing the project and linking to external resources such as publications, code repositories, and related websites.

Title	Lead principal investigator(s)	Lead program
A Data Management and Visualization Framework for Community Vulnerability to Hazards	Jeanne M. Jones and Kevin D. Henry	Western Geographic Science Center
A Web-Based Application for the Management and Visualization of Land-Use Scenario Data	Jason T. Sherba and Benjamin M. Sleeter	Western Geologic Science Center
Birds and the Bakken: Integration of Oil Well, Land Cover, and Species Distribution Data to Inform Conservation in Areas of Energy Development	Todd M. Preston and Rachel T. Bolus	Northern Rocky Mountain Science Center
Crowd-Sourced Earthquake Detections Integrated into Seismic Processing	Michelle R. Guy and Paul S. Earle	Geologic Hazards Team
Data Management Training Clearinghouse	John C. Nelson, Nancy J. Hoebelheinrich, and Tamar Norkin	Upper Midwest Environmental Sciences Center
Developing a USGS Legacy Data Inventory to Preserve and Release Historical USGS Data	John Faundeen and A. Lance Everette	Earth Resources Observation and Science Center
Development of Recommended Practices and Workflow for Publishing Digital Data through ScienceBase for Dynamic Visualization	Katherine J. Chase, Andrew R. Bock, and Roy Sando	Montana Water Science Center
Evaluating a New Open Source, Standards-Based Framework for Web Portal Development in the Geosciences	Richard P. Signell	Woods Hole Coastal & Marine Science Center
Facilitating the USGS Scientific Data Management Foundation by Integrating the Process into Current Scientific Workflow Systems	Colin B. Talbert, Drew A. Ignizio, Catherine Jarnevich, and Jeffrey T. Morisette	Fort Collins Science Center
Hunting Invasive Species with HTCondor: High Throughput Computing for Big Data and Next Generation Sequencing	S. Grace McCalla, Michael Fienen, Richard Erickson, Randall Hunt, and Jon Amberg	Upper Midwest Environmental Sciences Center
Integration of National Soil and Wetland Datasets: A Toolkit for Reproducible Calculation and Quality Assessment of Imputed Wetland Soil Properties	Eric T. Sundquist, Norman Bliss, Rusty Griffin, Sharon Waltman, and Lisamarie Windham-Myers	Branch of Regional Research
Integration of Phenological Forecast Maps for Assessment of Biodiversity: An Enterprise Workflow	Jake F. Weltzin, Theresa M. Crimmins, Alyssa Rosemartin, R. Lee Marsh, R. Sky Bristol, and Tim Kern	National Phenology Network
National Stream Summarization: Standardizing Stream-Landscape Summaries	Daniel J. Wieferich, Dana M. Infante, Marc Weber, Scott Leibowitz, Jeff Falgout, and Brad Williams	Core Science, Analytics, Synthesis, and Libraries

Community for Data Integration Projects

The FY 2016 projects represented many elements of the SSF, including data, information, communities of practice, applications, web services, data management, processing, analysis, preservation, and publishing and sharing. Many of the projects in FY 2016 focused on exploring distribution and visualization techniques for USGS data. A number of projects also focused on bringing various communities together to standardize processes and limit duplication of effort. Project PIs represented five out of the seven USGS mission areas: Climate and Land Use Change, Core Science Systems, Ecosystems, Natural Hazards, and Water. Each of the FY 2016 projects is described in detail below with references to completed products and deliverables. Many project teams continued working on deliverables after the end of the fiscal year. For example, journal articles and open-file reports associated with projects may take 6 to 12 months after the completion of the project to be published. Updates and additions to project accomplishments and deliverables will be made to the projects' records in ScienceBase (Community for Data Integration, 2016). Project titles in table 10 are hyperlinked to the ScienceBase record for each project, which provides links to related external resources such as publications, code repositories, and websites.

A Data Management and Visualization Framework for Community Vulnerability to Hazards

Lead PIs: Jeanne M. Jones and Kevin D. Henry

USGS research in the Western Geographic Science Center has produced several geospatial datasets estimating the time required to evacuate on foot from a Cascadia subduction zone earthquake-generated tsunami in the U.S. Pacific Northwest. These data, created as a result of research performed under the Risk and Vulnerability to Natural Hazards project, are useful for emergency managers and community planners but are not in the best format to serve their needs. This project explored options for formatting and publishing the data for consumption by external partner agencies and the general public.

The project team chose ScienceBase as the publishing platform, both for its ability to convert spatial data into web services and for its designation as an official platform for the public release of USGS data. Because the travel time map datasets are large vector files, the team planned to experiment with different extents, projections, and data resolutions to determine usability for internal use and for use by external partners. For the general public who reside in or near areas exposed to tsunamis and who are not likely to use web services, the project team experimented with Cesium, an innovative new JavaScript 3D mapping package, to create a web mapping application for exploration and visualization of the data.

Accomplishments

The accomplishments for this project are described in the following sections.

External Partner Needs Assessment

The original travel time maps were created as continuous value rasters and, for project purposes, were binned into 1-minute increments and converted to vector format. The project team uploaded these vector shapefiles to ScienceBase, created web feature services, and then contacted colleagues at the California Earthquake Clearinghouse to solicit feedback on the usefulness of the data. This external partner evaluated the product and made the following observations.

- The use of web services in ArcMap requires the Interoperability Extension, which many partners would not have.
- The 1-minute resolution time map is useful when zoomed in but is slow to load, so multiple services with times dissolved to 5- and 10-minute increments in a scale-dependent map would be ideal.
- The web feature service is nice for querying the map but is a slow format, and some users will prefer a simple image or tile service to use as an overlay in the field.
- If an image or map service is used that cannot be queried, create an accompanying semi-transparent layer with time labels.
- Serving the travel time maps by county is a good choice because many partners work at this level.
- Create an inventory of the web services with links and make this available to external partners so they know how to get to the data.

Datasets and Metadata in ScienceBase

The project team created metadata and uploaded eight travel time map vector shapefiles to ScienceBase (Wood and Jones, 2017). The maps are at the county level for Del Norte and Humboldt Counties in northern California, and the project team left them in the original 1-minute increments of travel time to safety. The team created web feature services in ScienceBase, loaded the services into the ArcMap desktop application, and tested the ability to query the maps for travel times at various locations. The team initially projected the maps to Web Mercator so they would work well with a commercial web base map, but they ended up using World Geodetic System (WGS) 84 to meet the requirement of their Cesium mapping application. The use of the ArcMap Interoperability Extension to access the services and the initial load time is a deterrent to use, but once in place, the services work well. Though the team appreciated the suggestions of the external partner, they realized that the creation of maps and metadata for so many different representations of the information was beyond the scope of this project.

Mapping Application Testbed

The project team decided that a 3D viewing platform would be the most appropriate platform for this project because the extra dimension of height is a critical aspect of tsunami risk and evacuation. Viewing in 3D allows the user to identify the role elevation plays in tsunami risk, as well as evacuation potential. The team wanted to impart this information to the public. Few browser-based 3D mapping libraries exist, but an emerging JavaScript library, Cesium, presented an opportunity to satisfy the requirements and also explore emerging technologies in visualization. The capabilities of this software were tested, allowing the team to identify best practices, which are outlined below, for using this software to present hazard exposure information.

- Input data *must* be projected to WGS 84.
- Datasets with a high number of features (over 10,000) or large file sizes *can* be displayed, though the browser's memory usage increases heavily.
- The preferred input data format for all but the largest datasets is topoJSON, providing seamless integration into the Cesium viewer and the ability to access data attributes and alter styling.
- By default, polygons are loaded into the 3D scene and projected flat against the spheroid of the Earth, but by changing a setting, the polygons can be "draped" over real-world terrain. This "draping" improves appearance but reduces performance, interfering with the ability to pan and interact smoothly with a map containing a large dataset.

Overall, Cesium had many promising features and may prove to be a viable alternative to the traditional web map (figure 2).

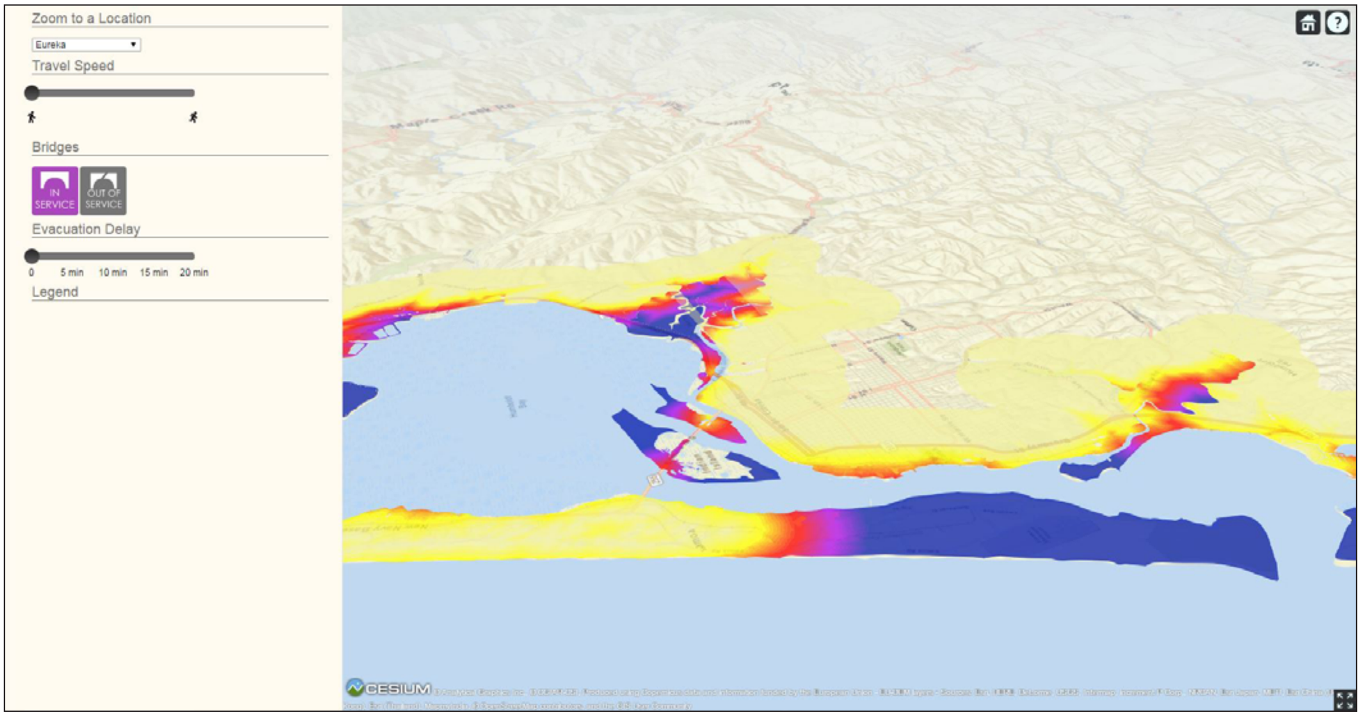

Zoom to a Location
Eureka
Travel Speed
Bridges
Evacuation Delay
Legend

Figure 2. Screenshot of the tsunami travel time map web application, displaying the map for Humboldt County, California, and the city of Eureka for a slow walking speed and bridges intact.

A Web-Based Application for the Management and Visualization of Land-Use Scenario Data

Lead PIs: Jason T. Sherba and Benjamin M. Sleeter

Land-use researchers need the ability to rapidly compare multiple land-use scenarios over a range of spatial and temporal scales and to visualize spatial and nonspatial data; however, land-use datasets are often distributed in the form of large tabular files and spatial files. These formats are not ideal for the way land-use researchers interact with and share these datasets. The size of these land-use datasets can quickly balloon in size. For example, land-use simulations for the Pacific Northwest, at 1-kilometer resolution, across 20 Monte Carlo realizations, can produce over 17,000 tabular and spatial outputs. A more robust management strategy is to store scenario-based, land-use datasets within a generalized land-use database designed specifically for managing a range of spatial and nonspatial land-use datasets. This type of database facilitates access to, and comparisons among, a range of scenario datasets as well as the generation and storage of metadata.

Methods for sharing land-use datasets with the public, in a way that encourages interaction with and understanding of land-use scenario data, are also lacking. Engaging and intuitive visualization capabilities are needed to effectively introduce non-scientists to the concept of scenarios and to successfully fulfill the USGS mandate of sharing datasets and science with the public.

This project aimed to facilitate management, sharing, and visualization of land change datasets with the creation of a database-driven web application. In collaboration with the University of California (UC) Berkeley Geospatial Innovation Facility (GIF) and ApexRMS, a web application was developed for visualizing and sharing multiple land-use scenario datasets. The application was built on top of a generalized land-use scenario database, providing a complete package for scenario management and visualization that can easily be applied to a range of scenario datasets. The generalized land-use database, which supports the preservation and sharing of simulation model inputs, outputs, and metadata, can provide a much needed data management solution for land-use researchers. The web viewer will allow researchers to quickly share land-use model results and associated datasets with colleagues and the public. Together these tools will serve as a complete package for managing and visualizing land-use datasets and as a framework for managing a range of other scenario-based datasets such as carbon stock and carbon flow datasets.

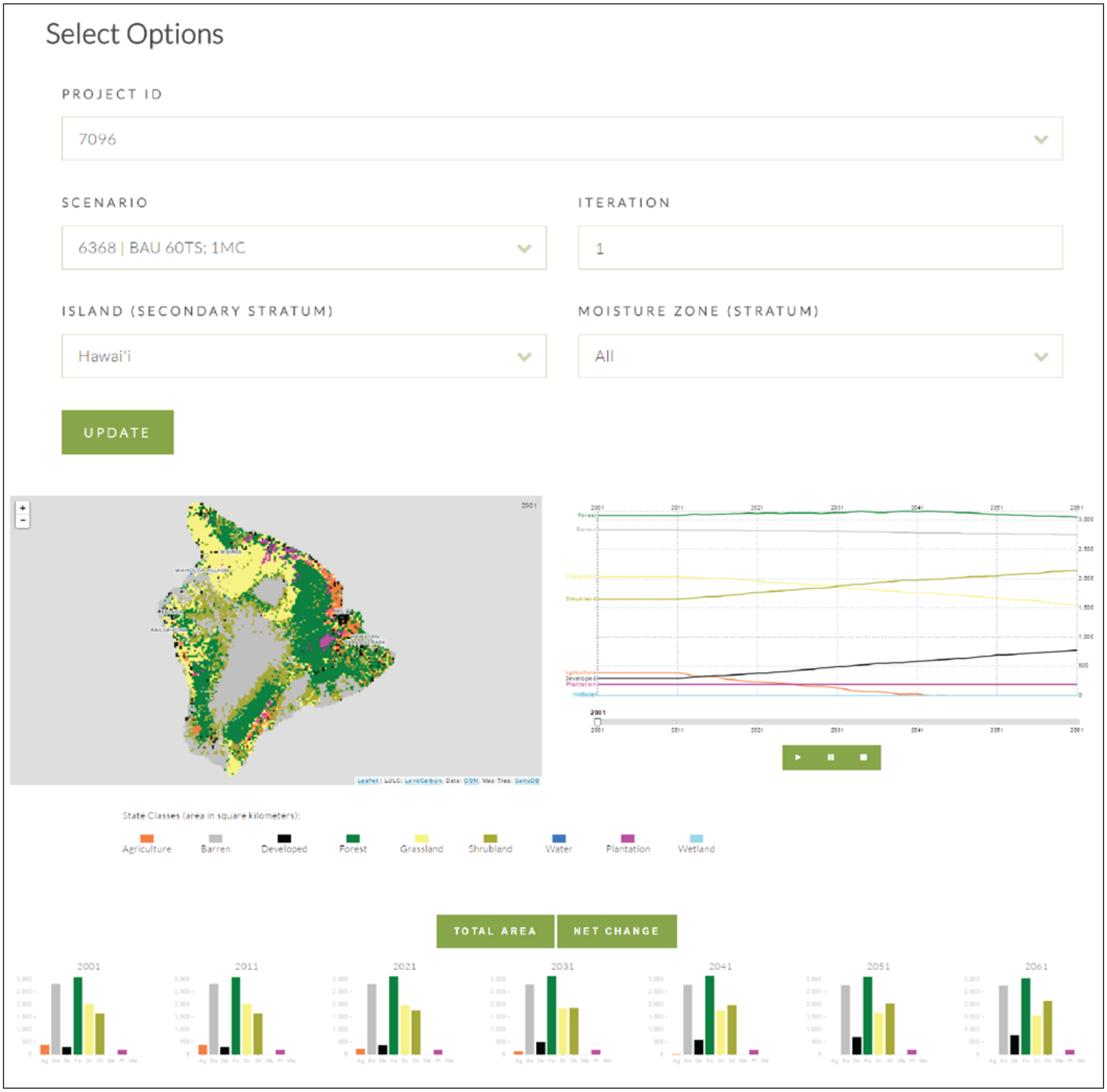
Accomplishments

The accomplishments for this project are described below.

- A limited Python-based API was developed to query and summarize results stored within a .ssim database. This API is a component of the web application and can also be used independently to manage data in a .ssim database. Code and documentation are available on GitHub: https://github.com/usgs/ssim-api.

- A method for adding external datasets to a .ssim database was added to the ST-Sim software package. ST-Sim is freely available at http://www.apexrms.com/state-and-transition-simulation-models/. This new tool allows users to add land-use scenario datasets from outside sources so that they can be stored, managed, and served from a .ssim database.

- The UC Berkeley GIF led the development of a web application for sharing and serving land-use and carbon scenario data managed by a .ssim database. A prototype was developed using results from a land-use and carbon model for the Hawaiian Islands. The beta version of web viewer is available at http://beta.landcarbon.org/ (fig. 3). The prototype also includes an admin console, for application deployment, that allows users to choose a .ssim database to connect to and specify scenarios for visualizing.



Select Options

PROJECT ID: 7096

SCENARIO: 6368 | BAU 60TS; 1MC

ITERATION: 1

ISLAND (SECONDARY STRATUM): Hawai'i

MOISTURE ZONE (STRATUM): All

State Classes (area in square kilometers):

🟧 Agriculture 🟩 Barren ⬛ Developed 🟩 Forest 🟨 Grassland 🟩 Shrubland 🟦 Water 🟪 Plantation 🟦 Wetland

Figure 3. This figure shows part of the web visualization dashboard from the Hawaiian Islands scenario viewer prototype: http://beta.landcarbon.org/. Users can select a scenario, iteration, and primary and secondary stratum and update results shown in the visualization dashboard.

Birds and the Bakken: Integration of Oil Well, Land Cover, and Species Distribution Data to Inform Conservation in Areas of Energy Development

Lead PIs: Todd M. Preston and Rachel T. Bolus

The goal of this project was to develop a novel methodology to combine the USGS Gap Analysis Program (GAP) national land cover and species distribution data with disturbance data to describe and predict how disturbance affects biodiversity. Specifically, the project team presented a case study examining how energy development in the Williston Basin can affect grassland birds; however, the methods developed are scalable and transferable to other types of habitat conversion (anthropogenic or natural), regions, and taxa. This project had six key components.

1. Develop a dataset delineating all oil well pads in the Williston Basin.
2. Develop a habitat conversion tool to determine the amount and previous land cover from spatially explicit disturbance.
3. Develop a species distribution tool that maps species richness for all input USGS GAP species distributions (two maps for the case study area: all avian species and grassland birds).
4. Develop an ecological effects tool that maps ecological effects (predation/nest success, non-native vegetation, avoidance, and noise in the case study) adjacent to spatially explicit disturbance data.
5. Develop a habitat suitability tool that integrates the outputs from the species distribution and ecological effects tools to determine habitat suitability following disturbance.
6. Develop models to predict habitat and biodiversity loss from predicted future disturbance.

It is the project team's expectation that the generated maps and models can assist resource managers by identifying sensitive, biodiverse areas.

Accomplishments

The project team made significant progress on the project components and are continuing to work on the project in FY 2017. The accomplishments for this project are described below.

- All oil well pads in the Williston Basin were delineated.
- All USGS GAP data were acquired and processed.
- Progress was made on the development of the required ArcGIS tools.
    - The habitat conversion description tool is 99 percent complete as of October 1, 2016. One small change is needed to improve graph aesthetics.
    - The species distribution map tool is 95 percent complete as of October 1, 2016 (fig. 4). The project team still needs to fix one component that filters by season.
    - The ecological effects layer is 50–75 percent complete. The work flow has been developed; however, the project team still needs to implement the workflow in the code.
    - The habitat suitability maps are 50–75 percent complete. These maps require output from the ecological effects layer. Once that layer is coded, completion of these maps should only take a few days.
    - The predictive model has not been started as of October 1, 2016, but should only take a few hours to complete once the output from the habitat suitability map is available.
- The peer-reviewed journal article is 95 percent complete. Much of the article is written; however, the project team requires the output from the unfinished tools to finalize the report.

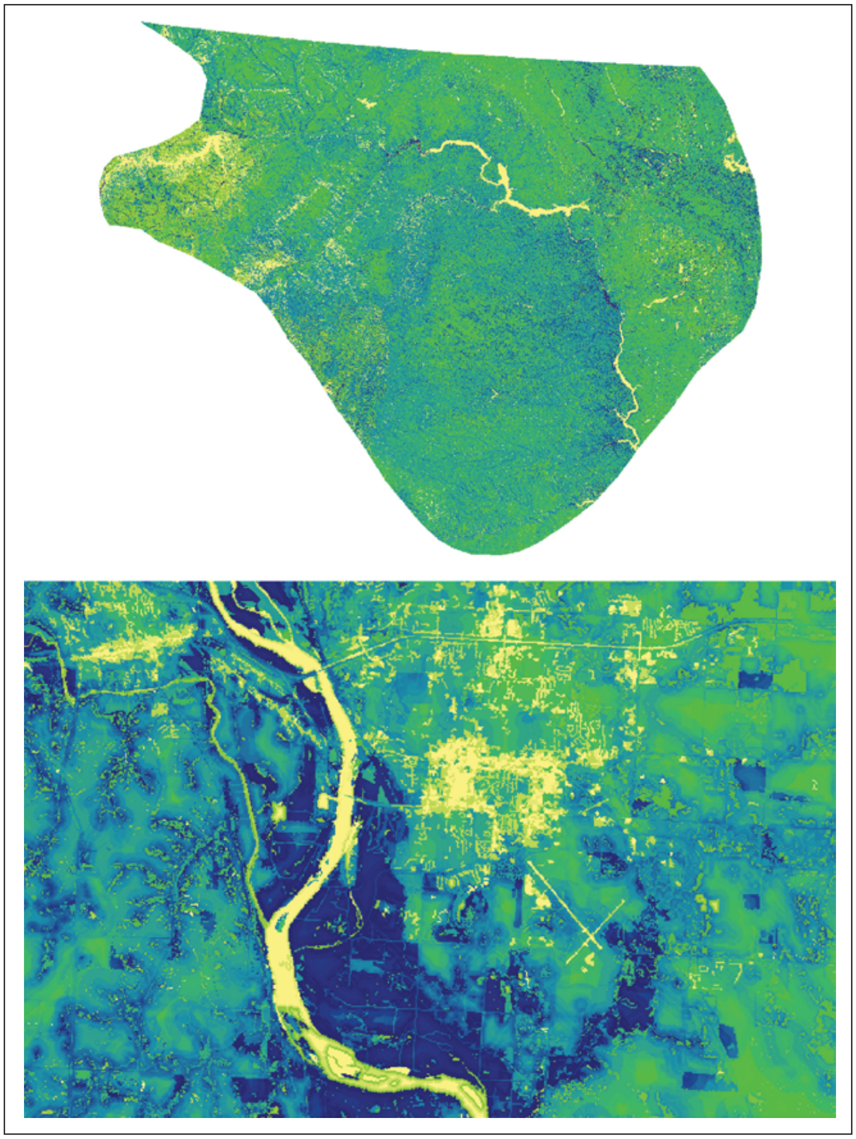


Figure 4. Output from the species distribution tool showing the avian species richness across the entire Williston Basin (top) and a close-up view along the Missouri River (bottom). Species richness increases from yellow to green to blue.

Crowd-Sourced Earthquake Detections Integrated into Seismic Processing

Lead PIs: Michelle R. Guy and Paul S. Earle

The goal of this project is to improve the USGS National Earthquake Information Center's (NEIC) earthquake detection capabilities through direct integration of crowd-sourced earthquake detections with traditional, instrument-based seismic processing. During the past 6 years, the NEIC has run a crowd-sourced system, called Tweet Earthquake Dispatch (TED), which rapidly detects earthquakes worldwide using data solely mined from Twitter messages, known as "tweets." The extensive spatial coverage and near instantaneous distribution of the tweets enable rapid detection of earthquakes often before seismic data are available in sparsely instrumented areas around the world. Although impressive for its speed, the tweet-based system has weaknesses, including missed events in non-populated areas, poor earthquake locations, and a 10-percent false trigger rate. To leverage the strengths and mitigate the weaknesses of both the crowd-sourced and instrument-based seismic systems, the project team used the rapid tweet-based detections as seeds for seismic processing of event location, phase data association, and magnitude determination. The rapid crowd-source detections allow the seismic systems to focus in on a region of interest, thus reducing the number of instrumental observations necessary to process an event and, in turn, accelerate the processing.

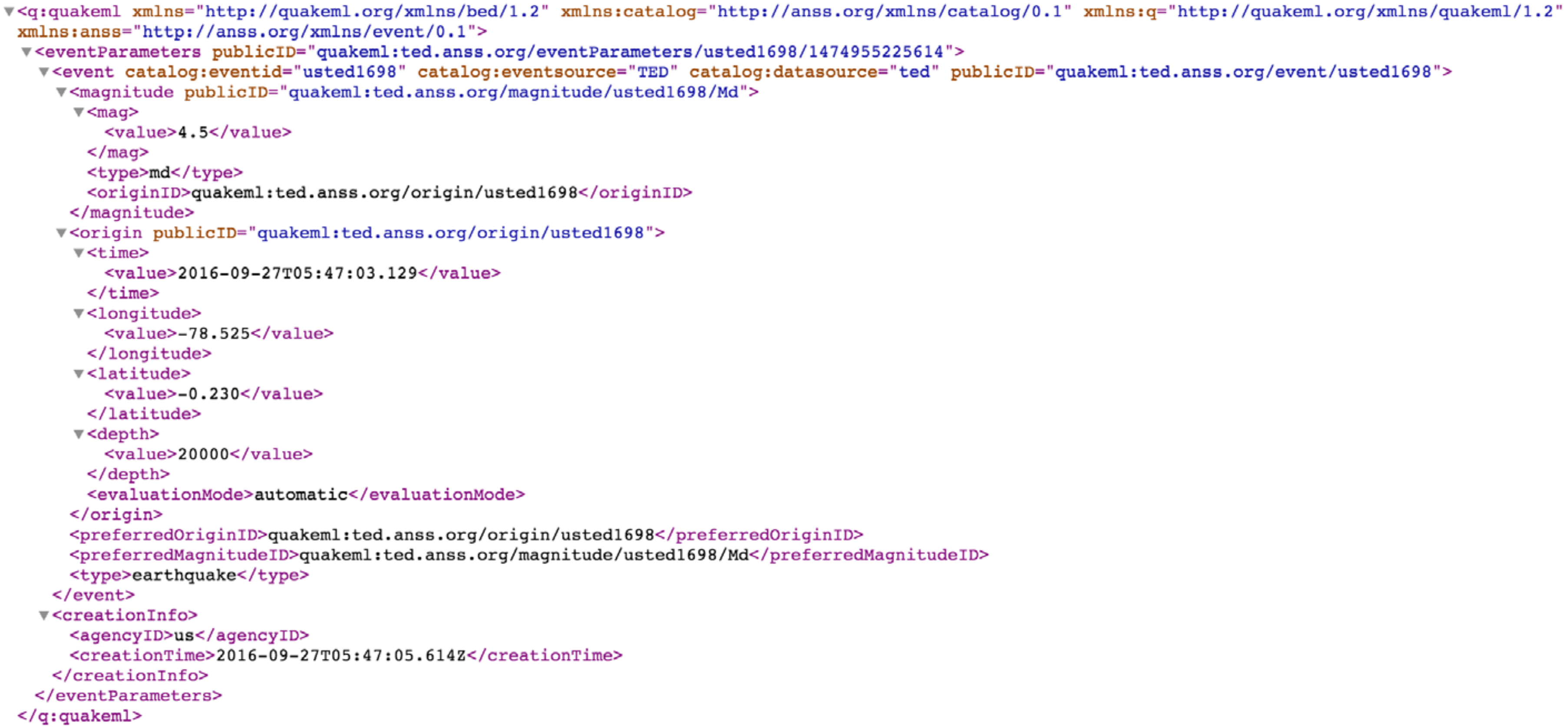
To seamlessly integrate these crowd-sourced detections with numerous existing processing systems, the detections were converted to an internationally recognized format for seismic data exchange and were distributed via existing standard mechanisms. Algorithmic improvements were made to the core tweet-based system to provide improved locations to better support the integration of the data. After successful integration of the Twitter and seismic data, the project team integrated an additional type of crowd-sourced earthquake detections that are derived from analysis of internet traffic and produced by the European-Mediterranean Seismological Centre (EMSC). These data, referred to as "flashsourcing," further increase the spatial coverage of the crowd-sourced detections.

Accomplishments

The accomplishments for this project are described in detail below.

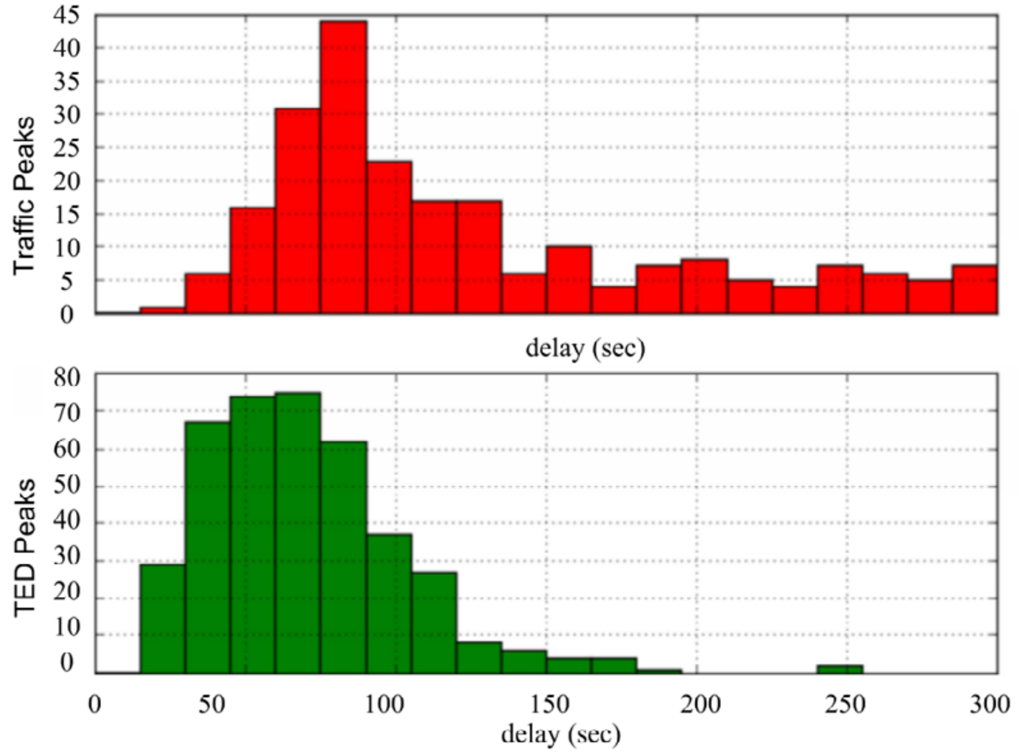
- The project team accomplished a number of tasks to support data sharing and integration.
  - They created, and now maintain, well-formatted, real-time, crowd-sourced earthquake detection data products, in an international standard seismic data exchange format, which seamlessly integrate with existing data distribution mechanisms and multiple data consumer applications for data sharing. An example of these QuakeML-formatted tweet-based earthquake detections is provided in figure 5. The corresponding code that produces these formatted detections is available at https://my.usgs.gov/bitbucket/projects/NEIC/repos/ted2quakeml/browse.
  - The project team created a metadata record for this dataset, which is available at https://www.sciencebase.gov/catalog/item/580108c3e4b0824b2d18bbd3.
  - The team now integrates, in real time, these well-formatted, tweet-based earthquake detections to the NEIC seismic processing system.
  - The team also provides tweet-based earthquake detections to EMSC in real time and receives EMSC-flashsourced earthquake detections in real time.
- The project team upgraded the search index and visual analysis interface to Elasticsearch v1.7 and Kibana v4.1. All updated code is available at https://my.usgs.gov/bitbucket/projects/NEIC/repos/ted2quakeml/browse.
- The project team made a containerized version of the application and successfully got it running at EMSC in France.

Current analysis results for integrating crowd-sourced detections with seismic systems show TED and Flashsourcing (peaks in web traffic from EMSC) systems detected felt earthquakes before enough seismic data were available for detection in 95 percent of the cases (figure 6). Due to deriving on the order of three felt detections daily and the timing of when the real-time data integration was established, there have not been enough statistically significant events for detailed analysis of how the seismic system is performing with the crowd-sourced detections as rapid inputs of possible earthquakes. Now that the real-time data integration is fully established, analysis will continue well beyond the life cycle of this CDI funded project.



```
▼<q:quakeml xmlns="http://quakeml.org/xmlns/bed/1.2" xmlns:catalog="http://anss.org/xmlns/catalog/0.1" xmlns:q="http://quakeml.org/xmlns/quakeml/1.2"
 xmlns:anss="http://anss.org/xmlns/event/0.1">
  ▼<eventParameters publicID="quakeml:ted.anss.org/eventParameters/usted1698/1474955225614">
    ▼<event catalog:eventid="usted1698" catalog:eventsource="TED" catalog:datasource="ted" publicID="quakeml:ted.anss.org/event/usted1698">
      ▼<magnitude publicID="quakeml:ted.anss.org/magnitude/usted1698/Md">
        ▼<mag>
           <value>4.5</value>
         </mag>
         <type>md</type>
         <originID>quakeml:ted.anss.org/origin/usted1698</originID>
       </magnitude>
      ▼<origin publicID="quakeml:ted.anss.org/origin/usted1698">
        ▼<time>
           <value>2016-09-27T05:47:03.129</value>
         </time>
        ▼<longitude>
           <value>-78.525</value>
         </longitude>
        ▼<latitude>
           <value>-0.230</value>
         </latitude>
        ▼<depth>
           <value>20000</value>
         </depth>
         <evaluationMode>automatic</evaluationMode>
       </origin>
       <preferredOriginID>quakeml:ted.anss.org/origin/usted1698</preferredOriginID>
       <preferredMagnitudeID>quakeml:ted.anss.org/magnitude/usted1698/Md</preferredMagnitudeID>
       <type>earthquake</type>
     </event>
    ▼<creationInfo>
       <agencyID>us</agencyID>
       <creationTime>2016-09-27T05:47:05.614Z</creationTime>
     </creationInfo>
   </eventParameters>
 </q:quakeml>
```

Figure 5. Example of well-formatted, real-time, crowd-sourced earthquake detection data products, in international standard seismic data exchange format, QuakekML, from Tweet Earthquake Dispatch (TED) detection. Metadata for these data available at https://www.sciencebase.gov/catalog/item/580108c3e4b0824b2d18bbd3 (ScienceBase web page). This example is for a confirmed M2.9 NE of Quito Ecuador (Earthquake Hazards Program, 2016).



Crowd-sourced detection method	Number of detections	Minimum delay (s)	Median delay (s)
EMSC Flashsourcing (red)	257	29	109
TED (green)	398	16	65

Figure 6. Distribution of the time delays to detect a felt earthquake for European-Mediterranean Seismological Centre (EMSC) Flashsourcing (red) and Tweet Earthquake Dispatch (TED) (green). (s, sec, second)

Data Management Training Clearinghouse

Lead PIs: John C. Nelson, Nancy J. Hoebelheinrich, and Tamar Norkin

The purpose of the Data Management Training (DMT) Clearinghouse project was twofold. First, the project aimed to increase discoverability and accessibility of the wealth of learning resources that have been developed to inform and train scientists about data management in the Earth sciences. Secondly, the project team wanted to facilitate the use of these learning resources by providing descriptive information (metadata) that can help research scientists, students, or teachers assess whether the resource would be appropriate and useful for their needs.

The project team established the following objectives for the project.

1. Create an online, searchable, and browsable clearinghouse of learning resources on data management in the Earth sciences.

2. Provide a mechanism for creators or managers of learning resources to easily submit information about their resources into the clearinghouse.

3. Provide the means for content submitters to describe the learning resources using a description schema that can increase discoverability and describe the contextual framework for which they were created.

4. Provide a metadata submission form with online help guides available to assist content submitters in contributing complete and accurate information about their learning resources.

5. Develop a basic crowd-sourced, quality-assured mechanism for building and sustaining the clearinghouse past the initial grant-funded development of the clearinghouse.

Accomplishments

The project team created the DMT Clearinghouse hosted within the web domain of ESIP, using the existing Drupal content management system environment at http://dmtclearinghouse.esipfed.org (fig. 7). The clearinghouse contains the following components.

- An online metadata registry with a publicly available, facet-searchable and browsable inventory of learning resources about data management practices (fig. 8). This registry includes references to community-developed frameworks for data management such as the USGS SSF and its Science Data Lifecycle Model.
- Home and About pages that provide context for the project, including goals, approach, and partner information.
- Help pages including an FAQ section and assistance with content submission, which will be iterated as more questions arise from user experience.
- A submission form for including metadata about the learning resources in the clearinghouse.
- A simple workflow process for content submission, review/editing, and release to publication.

The project team developed a submission form used to collect information about the learning resources that are to be included in the clearinghouse inventory. The submission form is available to users who have registered for a free ESIP login account. The submission form uses key metadata fields from the Learning Resource Metadata Initiative (LRMI) schema that are adapted and applied to the types of learning resources targeted for inclusion in the clearinghouse. The LRMI schema was developed by the education community, has been endorsed by Schema.org, and is maintained by DCMI.org, the Dublin Core Metadata Initiative. The submission form provides a multi-level interface of required and recommended metadata fields for ease of use by content submitters and reviewers/editors. The form is also populated with field definitions, tips, techniques, and examples of metadata values to facilitate content submission. The submission form underwent testing for usability by members of each of the initial partners (USGS, DataONE, and ESIP Federation) and has been improved per initial user feedback. The form was used successfully to create an initial inventory of learning resources for the clearinghouse, which was launched in October 2016.

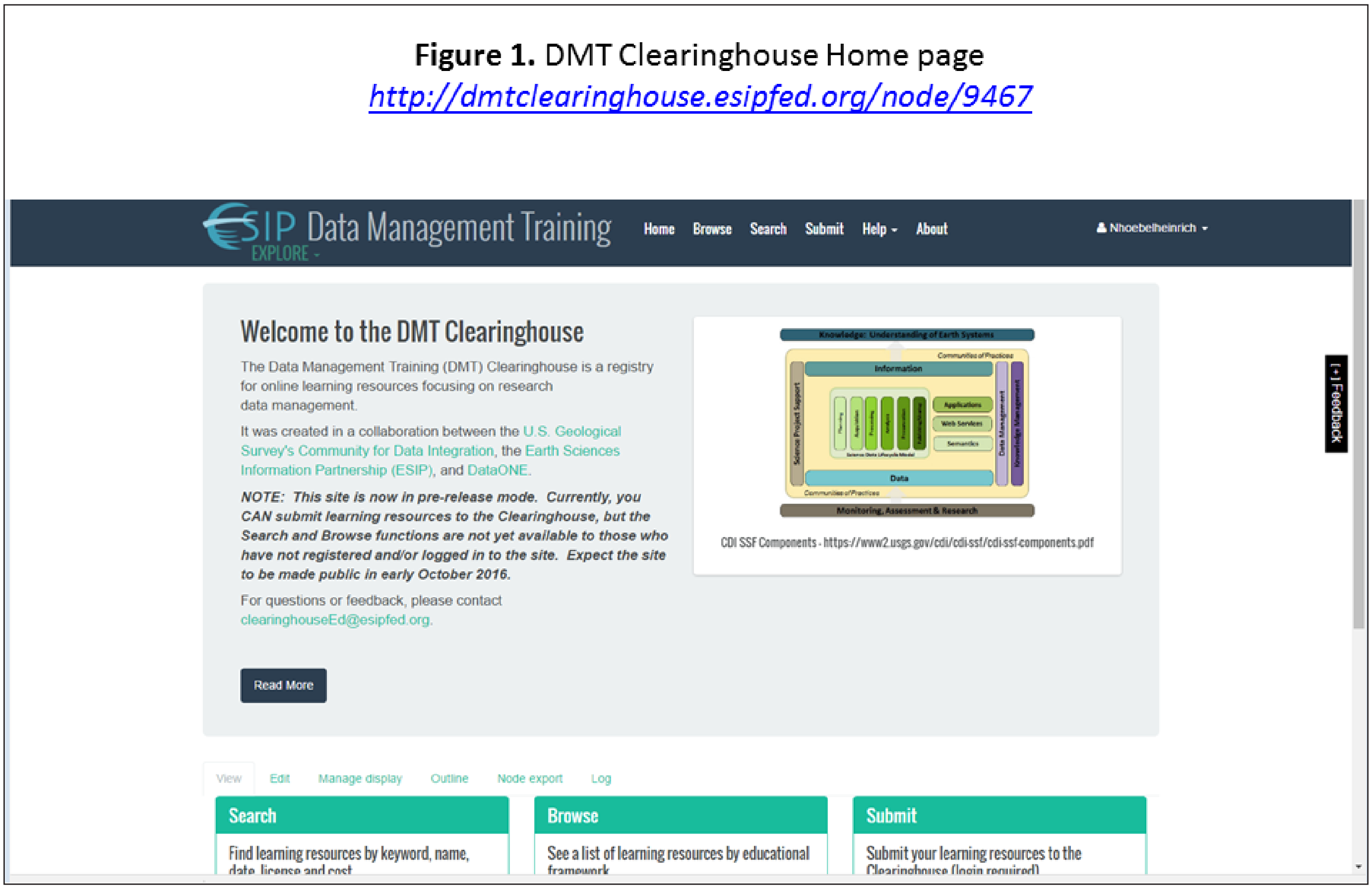


Figure 1. DMT Clearinghouse Home page
http://dmtclearinghouse.esipfed.org/node/9467

Welcome to the DMT Clearinghouse

The Data Management Training (DMT) Clearinghouse is a registry for online learning resources focusing on research data management.

It was created in a collaboration between the U.S. Geological Survey's Community for Data Integration, the Earth Sciences Information Partnership (ESIP), and DataONE.

NOTE: This site is now in pre-release mode. Currently, you CAN submit learning resources to the Clearinghouse, but the Search and Browse functions are not yet available to those who have not registered and/or logged in to the site. Expect the site to be made public in early October 2016.

For questions or feedback, please contact clearinghouseEd@esipfed.org.

Read More

CDI SSF Components - https://www2.usgs.gov/cdi/cdi-ssf/cdi-ssf-components.pdf

Search

Find learning resources by keyword, name, date, license and cost

Browse

See a list of learning resources by educational framework

Submit

Submit your learning resources to the Clearinghouse (login required)

Figure 7. Screenshot of the Data Management Training Clearinghouse web page at http://dmtclearinghouse.esipfed.org/node/9467.

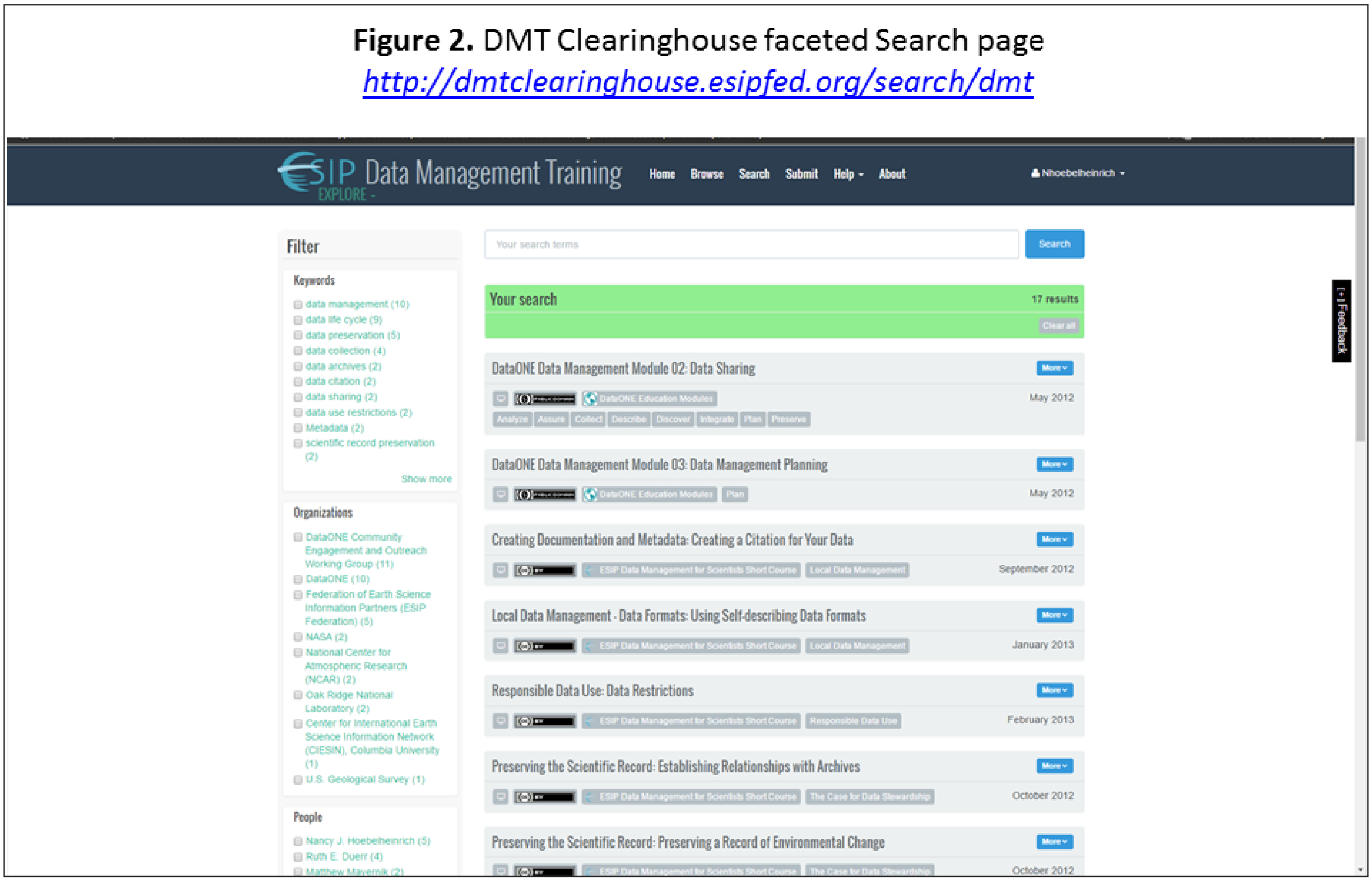


Figure 2. DMT Clearinghouse faceted Search page
http://dmtclearinghouse.esipfed.org/search/dmt

Filter

Keywords

- □ data management (10)
- □ data life cycle (9)
- □ data preservation (5)
- □ data collection (4)
- □ data archives (2)
- □ data citation (2)
- □ data sharing (2)
- □ data use restrictions (2)
- □ Metadata (2)
- □ scientific record preservation (2)

Organizations

- □ DataONE Community Engagement and Outreach Working Group (11)
- □ DataONE (10)
- □ Federation of Earth Science Information Partners (ESIP Federation) (5)
- □ NASA (2)
- □ National Center for Atmospheric Research (NCAR) (2)
- □ Oak Ridge National Laboratory (2)
- □ Center for International Earth Science Information Network (CIESIN), Columbia University (1)
- □ U.S. Geological Survey (1)

People

- □ Nancy J. Hoebelheinrich (5)
- □ Ruth E. Duerr (4)
- □ Matthew Mayernik (2)

Your search terms

Your search — 17 results

DataONE Data Management Module 02: Data Sharing
DataONE Education Modules — Analyze, Assure, Collect, Describe, Discover, Integrate, Plan, Preserve — May 2012

DataONE Data Management Module 03: Data Management Planning
DataONE Education Modules — Plan — May 2012

Creating Documentation and Metadata: Creating a Citation for Your Data
ESIP Data Management for Scientists Short Course — Local Data Management — September 2012

Local Data Management - Data Formats: Using Self-describing Data Formats
ESIP Data Management for Scientists Short Course — Local Data Management — January 2013

Responsible Data Use: Data Restrictions
ESIP Data Management for Scientists Short Course — Responsible Data Use — February 2013

Preserving the Scientific Record: Establishing Relationships with Archives
ESIP Data Management for Scientists Short Course — The Case for Data Stewardship — October 2012

Preserving the Scientific Record: Preserving a Record of Environmental Change
ESIP Data Management for Scientists Short Course — The Case for Data Stewardship — October 2012

Figure 8. Screenshot of the Data Management Training Clearinghouse faceted search page at http://dmtclearinghouse.esipfed.org/search/dmt.

Throughout the project, the team introduced the DMT Clearinghouse concept and product at various public venues including Research Data Access and Preservation Summit 2016 (poster presentation), DataONE User's Group meeting in July 2016 (presentation and user feedback), ESIP 2016 Summer Meeting (presentation and user feedback), DataONE All Hands meeting in September 2016 (user testing and content submission), and the ESIP Data Stewardship Committee's October 2016 monthly meeting (website launch presentation). The DMT Clearinghouse has generated interest from a number of other Earth-science-affiliated groups such as the National Ecological Observatory Network, the National Center for Ecological Analysis and Synthesis, Data Carpentry, Research Data Alliance (U.S. Education Interest Group), and the Federal Scientific and Technical Information Managers Group. These groups are interested in submitting content to the DMT Clearinghouse and collaborating on activities such as taxonomy development, development of core skills for data professionals, and identifying gaps in topic coverage for data management learning resources.

The project team has identified future opportunities for more user testing, content submission workshop events, and marketing and outreach to build and sustain the clearinghouse. The team has also identified future enhancements for the clearinghouse.

- Development of robust workflows to facilitate crowdsourcing of content submission and efficient review and publication of submitted content for clearinghouse reviewers and editors.
- Implementation of automatic link checking to ensure the currency of the links to the external locations of the learning resources.
- Collection of analytics related to page hits on the clearinghouse and individual learning resources, especially in light of the use of the Schema.org-endorsed LRMI metadata scheme.
- Automatic harvesting of metadata from other data lesson catalogs or sources.
- Development of capabilities to push information about the DMT Clearinghouse learning resources out via Drupal's learning registry module.

Developing a USGS Legacy Data Inventory to Preserve and Release Historical USGS Data

Lead PIs: John L. Faundeen and Anthony L. Everette

As one of the largest and oldest Earth science organizations in the world, the legacy of the USGS is its data and the scientific knowledge derived from them. It is widely understood, however, that high-quality data collected and analyzed during long completed projects are stored in case files, file cabinets, and hard drives across USGS. Despite their potential significance to current USGS mission and program objectives, these "legacy data" are unavailable to the scientific community, and the challenges of preserving them increase as technologies change and the data producers retire.

Fortunately, the USGS has a long history of proactively researching and developing solutions to data management needs and is a lead agency in establishing meaningful and actionable policies that facilitate public data release to the scientific community. In recent years, several USGS projects have investigated tools and methods for preserving and publishing legacy data. For example, one of the most comprehensive Bureau-level legacy data preservation efforts to date was the USGS Data Rescue Program, which provided funding, tools, and support to preserve more than 100 legacy datasets determined to be at imminent risk of permanent loss or damage. Subsequent projects such as the Legacy Data Inventory and Reporting System Project and the North American Bat Data Recovery Project have investigated the application of technology to more effectively inventory, report, and distribute legacy data.

Based on these experiences, the project team has seen three primary challenges to legacy data preservation and release.

1. Inventory: What legacy data does the USGS have?
2. Evaluation and Prioritization: How should USGS scientists prioritize their legacy data inventories and determine which products to preserve first?
3. Funding and Support: How much time and effort are required to preserve and publish a specific legacy data product? What options are available for funding support?

To help the USGS and its science centers address these challenges, the "Developing a USGS Legacy Data Inventory to Preserve and Release Historical USGS Data" project had three primary objectives.

1. Create a USGS legacy data inventory that catalogs and describes known USGS legacy datasets.
2. Develop methods to evaluate and prioritize the legacy data inventory based on USGS mission objectives and imminent risks of loss or damage.
3. Preserve and release select, priority legacy datasets at risk of damage or loss, and document time and resource costs required to complete each phase of the data lifecycle model: plan, acquire, process, preserve, and publish.

Accomplishments

To begin the process of building a USGS legacy data inventory, the project team conducted a USGS-wide request for legacy data submission during April and May 2016. The team received 43 legacy data submissions from 20 USGS science centers involving all USGS mission areas. In addition, the unfunded Data Rescue Program requests received during 2006–2013 are being added to the USGS legacy data inventory along with 300+ legacy data records being migrated by the Fort Collins Science Center. The legacy data inventory is available at https://www.fort.usgs.gov/ldi/legacy-products.

During June and July 2016, the project team developed a method for legacy data submission reviewers to score the risk and significance factors associated with each legacy data product, which can be viewed at https://www.fort.usgs.gov/ldi/evaluations-reports/dar16. The team then developed simple algorithms that use the reviewers' scores to calculate inventory-level priorities. To help communicate legacy data priorities, they developed interactive reports for USGS mission areas, programs, and science centers.

Based on the risk and significance report, the team selected the top five legacy data products and are preserving and publishing these as official USGS data releases (table 11).

Preservation activities were underway as of October 1, 2016, for all five selected preservation projects and were on schedule for public release in 2017. Based on that work, the project team collected data on the resources required to complete each step of their data management plans (for example, acquire, process, analyze, publish) to better inform future legacy data preservation and release estimates.

Table 11. Five legacy data products being preserved and published through the U.S. Geological Survey (USGS) project "Developing a USGS Legacy Data Inventory to Preserve and Release Historical USGS Data."

Data product	Data time range	Data contact
Great Lakes Bathythermograph Data	1954	Sophia Dabrowski, Great Lakes Science Center
Friends Creek at Argenta Gage-Height Data	1971–1982	John Latour, Illinois Water Science Center
San Andreas Fault (Parkfield, CA) Magnetotelluric (MT) Survey Data	1990	Michaela Johnson, Crustal Geophysics and Geochemistry Science Center
Kanab Creek Repeat Photography Collection (figure 9)	1872–2010	Terry Arundel, Southwest Biological Science Center
Historical California River Channel Survey Data	1974–2013	Mary Ann Madej, Western Ecological Research Center

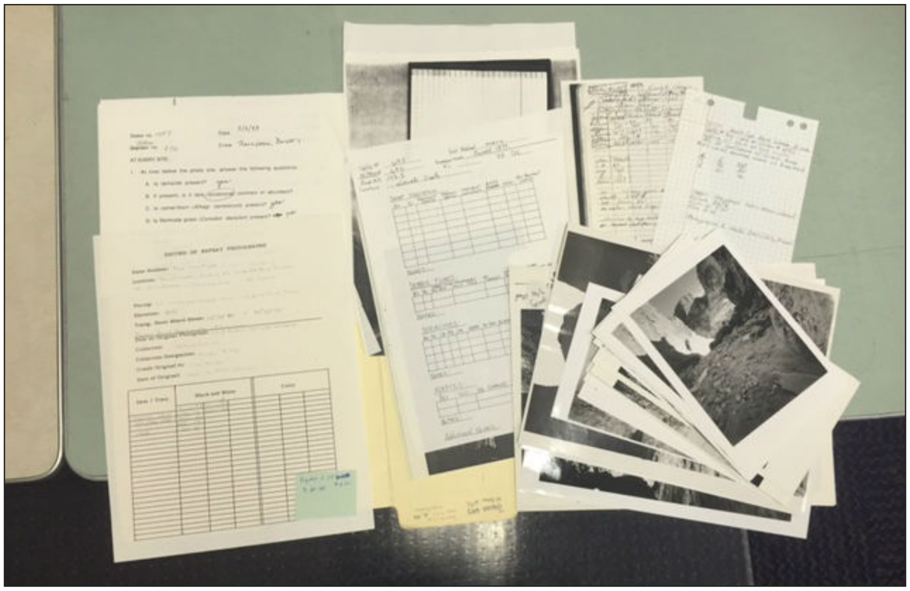


Figure 9. Repeat photography data such as the Kanab Creek repeat photo collection are organized by stake folder, which includes the stake location data, original and repeat photographers' image data, field notes, and film and print images—all of which contain components required to make a complete repeat photo record.

Development of Recommended Practices and Workflow for Publishing Digital Data through ScienceBase for Dynamic Visualization

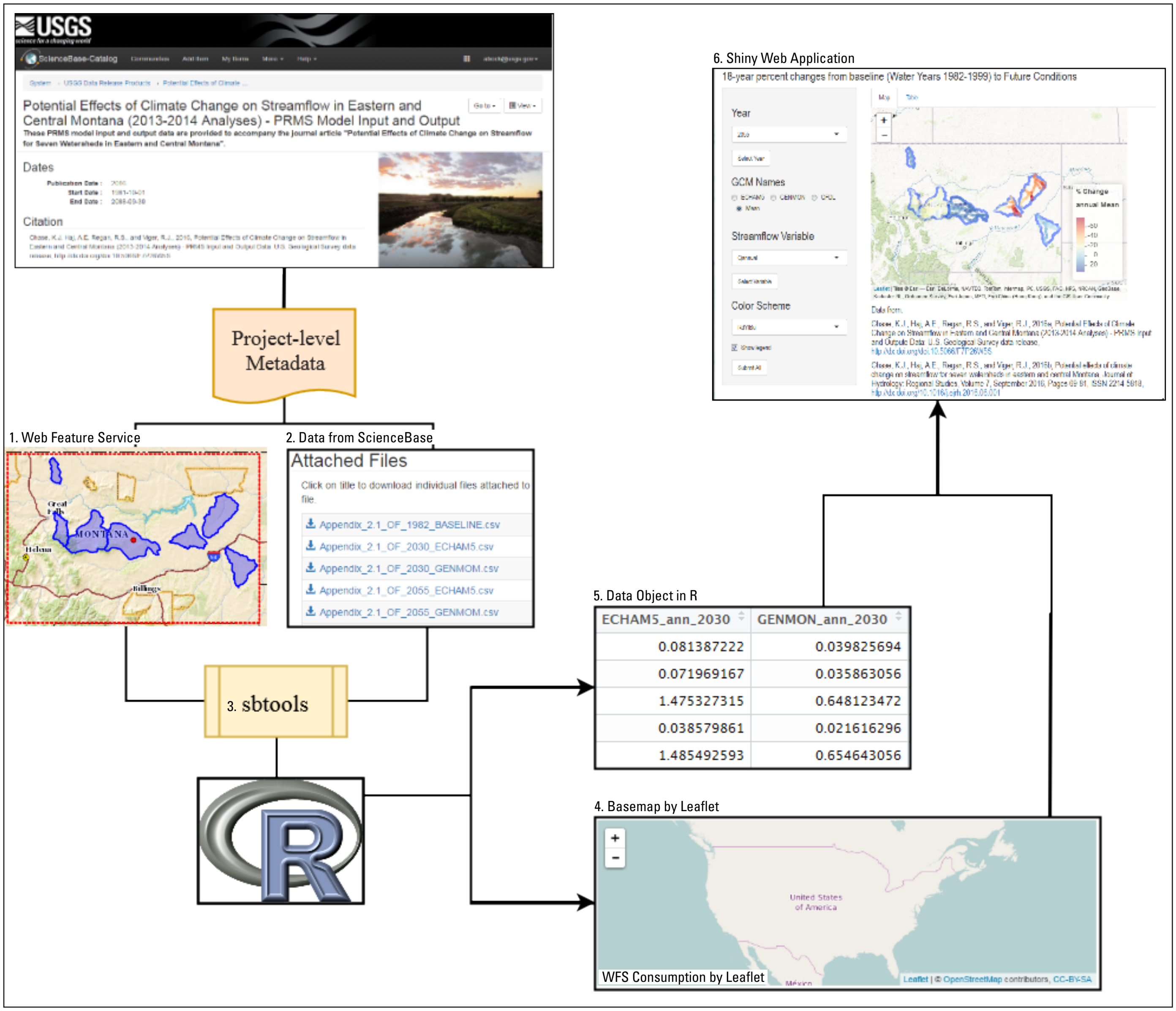
Lead PIs: Katherine J. Chase, Andrew R. Bock, and Roy Sando

The purpose of this project was to document processes for USGS scientists to organize and share data using ScienceBase and to provide an example interactive mapping application to display those data. Data and maps from Chase and others (2016a, b) were used for the example interactive maps.

Accomplishments

The accomplishments for this project are described below.

- The project team developed an interactive mapping application in R that connects to data on ScienceBase, using Shiny, Leaflet (Cheng and Xie, 2016), and sbtools (Winslow and others, 2016) (fig. 10). USGS scientists can refer to the R code in the mapping application to build their own interactive maps. Code is available at the USGS Bitbucket Repository (https://my.usgs.gov/bitbucket/projects/CDI/repos/interactive-mapping). The map application is a lightweight desktop



Potential Effects of Climate Change on Streamflow in Eastern and Central Montana (2013-2014 Analyses) - PRMS Model Input and Output

These PRMS model input and output data are provided to accompany the journal article "Potential Effects of Climate Change on Streamflow for Seven Watersheds in Eastern and Central Montana".

Dates

Publication Date : 2016
Start Date : 1981-10-01
End Date : 2099-09-30

Citation

Chase, K.J., Haj, A.E., Regan, R.S., and Viger, R.J., 2016, Potential Effects of Climate Change on Streamflow in Eastern and Central Montana (2013-2014 Analyses) - PRMS Input and Output Data: U.S. Geological Survey data release, http://dx.doi.org/10.5066/F7TP26N5

Attached Files

Click on title to download individual files attached to file.

- Appendix_2.1_OF_1982_BASELINE.csv
- Appendix_2.1_OF_2030_ECHAM5.csv
- Appendix_2.1_OF_2030_GENMOM.csv
- Appendix_2.1_OF_2055_ECHAM5.csv
- Appendix_2.1_OF_2055_GENMOM.csv

ECHAM5_ann_2030	GENMON_ann_2030
0.081387222	0.039825694
0.071969167	0.035863056
1.475327315	0.648123472
0.038579861	0.021616296
1.485492593	0.654643056

18-year percent changes from baseline (Water Years 1982-1999) to Future Conditions

Data from:

Chase, K.J., Haj, A.E., Regan, R.S., and Viger, R.J., 2016a, Potential Effects of Climate Change on Streamflow in Eastern and Central Montana (2013-2014 Analyses) - PRMS Input and Outpute Data: U.S. Geological Survey data release, http://dx.doi.org/doi:10.5066/F7TP26N5S

Chase, K.J., Haj, A.E., Regan, R.S., and Viger, R.J., 2016b, Potential effects of climate change on streamflow for seven watersheds in eastern and central Montana. Journal of Hydrology: Regional Studies, Volume 7, September 2016, Pages 69-81, ISSN 2214-5818, http://dx.doi.org/10.1016/j.ejrh.2016.06.001

Figure 10. Creation of an interactive map using a ScienceBase data release, R, Leaflet, and Shiny. From the ScienceBase data release, the (1) web feature service (WFS) containing the geographic entities and (2) data indexed to the geographic entities are brought into R environment through the use of the (3) sbtools package. The WFS components can be consumed through and added to (4) open-source base maps utilizing the Leaflet library through R, while data can be summarized and directly read as (5) data objects in R. The (6) Shiny web application framework is then used to build a user interface that connects the geographic entities to the underlying data objects summarized in R to allow a simple level of user interaction.

interactive map with two intended audiences: (1) USGS scientists interested in learning how to develop applications from a ScienceBase data release and (2) researchers, collaborators, and cooperators associated with regional research efforts related to the work of Chase and others (2016a). The four components to the application are (1) a user interface where the user selects the period of record, streamflow variable, and climate dataset of interest; (2) a map that displays the data based on the user's query from the first component for the stream features from Chase and others (2016a); (3) an interactive map on which users can query the hydrographic features associated with Chase and others (2016a, b); and (4) a dynamic table which displays data for the selected climate datasets across all periods of record and streamflow variables based on the hydrographic feature selected in the third component. The developers used R and its associated libraries based upon the support and expertise for R language and tools that are in-house at the Office of Water Information. The developers felt that, in the future, such applications should leverage as many existing USGS tools (such as the sbtools package in R) as possible.

- The project team documented the process to organize and share data on ScienceBase as well as the interactive map example in USGS Open-File Report 2016–1202 (Chase and others, 2017).
- The project team developed a webinar to share this process with USGS scientists, and the webinar took place on February 22, 2017.

Evaluating a New Open-Source, Standards-Based Framework for Web Portal Development in the Geosciences

Lead PI: Richard P. Signell

Web portals are one of the principal ways geospatial information can be communicated to the public. A few prominent USGS examples are the Geo Data Portal (http://cida.usgs.gov/gdp/ [URL is accessible with Google Chrome]), EarthExplorer (http://earthexplorer.usgs.gov/), the former Derived Downscaled Climate Projection Portal, the Alaska Portal Map (http://alaska.usgs.gov/portal/), the Coastal Change Hazards Portal (http://marine.usgs.gov/coastalchangehazardsportal/), and The National Map (http://nationalmap.gov/). Currently, web portals are developed at relatively high effort and cost, with web developers working with highly skilled data specialists on custom solutions that meet user needs. To address this issue, the Australian National Government funded the development of an open-source framework for building web portals called TerriaJS, which began in early 2015 (TerriaJS, 2015). TerriaJS takes advantages of capabilities in modern browsers to deliver a browser-only solution that consumes web map services from Esri and Open Geospatial Consortium (OGC), the most commonly used web map service (WMS) standards employed at the USGS and throughout the geoscience community. Because TerriaJS runs completely within the web browser, it is possible to generate custom portals by using simple configuration files that can be located anywhere on the internet. This means that basic portals based on WMSs can be created by non-JavaScript developers such as scientists, environmental managers, and emergency response support personnel. It also means they can be constructed rapidly, in hours or days instead of weeks or months. Finally, TerriaJS could also reduce the development cost of more sophisticated portals by providing a broad framework that covers a wide range of common portal mapping needs.

While TerriaJS appeared promising, we needed to more closely examine its capabilities, potential, and risks to fully understand its value. What are the generic portal needs that can be addressed by the existing framework? Does the framework architecture allow expansion to address more sophisticated needs beyond basic web service mapping? Does the framework have sufficient documentation and interaction or encouragement from the lead developers to actually function as a community-driven open-source project, or are enhancements only reasonably created by core developers? The project team planned to address these questions by taking a deep dive into TerriaJS; adding specific enhancements needed to support access to meteorologic, oceanographic, and hydrologic model data; and using the framework to create several web portals using a combination of developer resources from the USGS Office of Water Information and the Australian CSIRO/Data61 team.

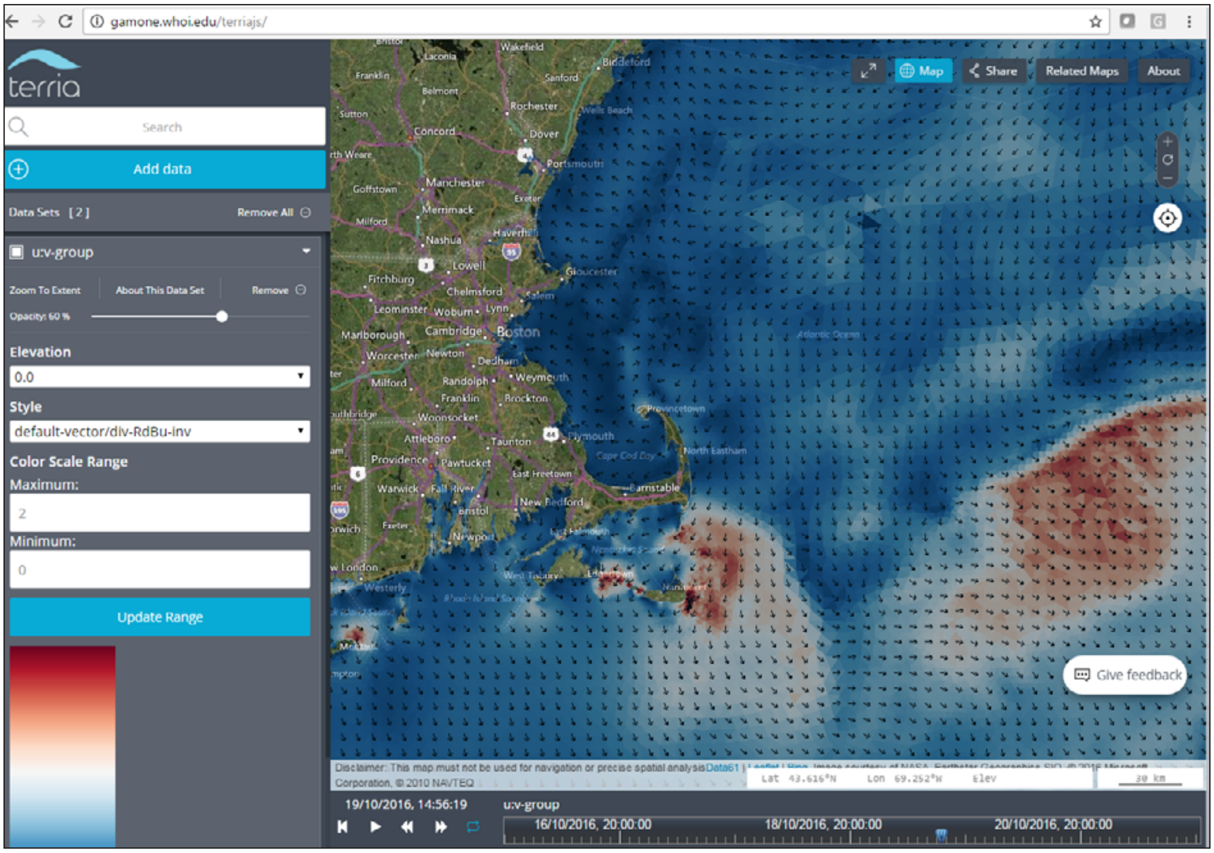
Accomplishments

The project team achieved the main objective of assessing the role of TerriaJS in the USGS suite of available tools for creating web services. Using a combination of USGS and Data61 developer resources, the team made critical enhancements for dealing with meteorologic, oceanographic, and hydrologic model data; tested deployment in the USGS computational environment; and developed several demonstrations using TerriaJS, which allowed them to assess performance and ease of installation and use.

The code enhancements enabled by the CDI project were merged back into the master branch on GitHub (TerriaJS, 2015). These enhancements were controls that allow the user to select layers, change color ranges, and select styles from ncWMS and ncWMS2 services integrated into THREDDS Data Servers (fig. 11). These enhancements allow effective exploration of model data via WMS in TerriaJS, and thanks to this project, this capability is now available not only to the USGS, but to everyone.

A configuration was set up to explore multiple time-dependent datasets from geospatial standards, purported strengths of TerriaJS. The project team created a bird migration example using Cesium Markup Language (CZML), a JavaScript Object Notation (JSON) that allows representation of geometric information in space and time, to display the bird migration data, and using WMS from a National Oceanic and Atmospheric Administration (NOAA) THREDDS server to display underlying monthly temperature data (fig. 12). This example portal is available at http://tinyurl.com/terriajs-birds. This example serves to show the power and flexibility of TerriaJS in handling time-dependent data, as well as ease of configuration by end users.

The USGS system for automatic deployment of services controlled by continuous integration on a code repository was also tested. The system performed as advertised; code deployed on a staging branch on Bitbucket, available at https://my.usgs.gov/bitbucket/projects/CDI/repos/terriamap/, automatically was deployed on the USGS endpoint.


gamone.whoi.edu/terriajs/

Data Sets [2]

u:v-group

Opacity: 60 %

Elevation: 0.0

Style: default-vector/div-RdBu-inv

Color Scale Range

Maximum: 2

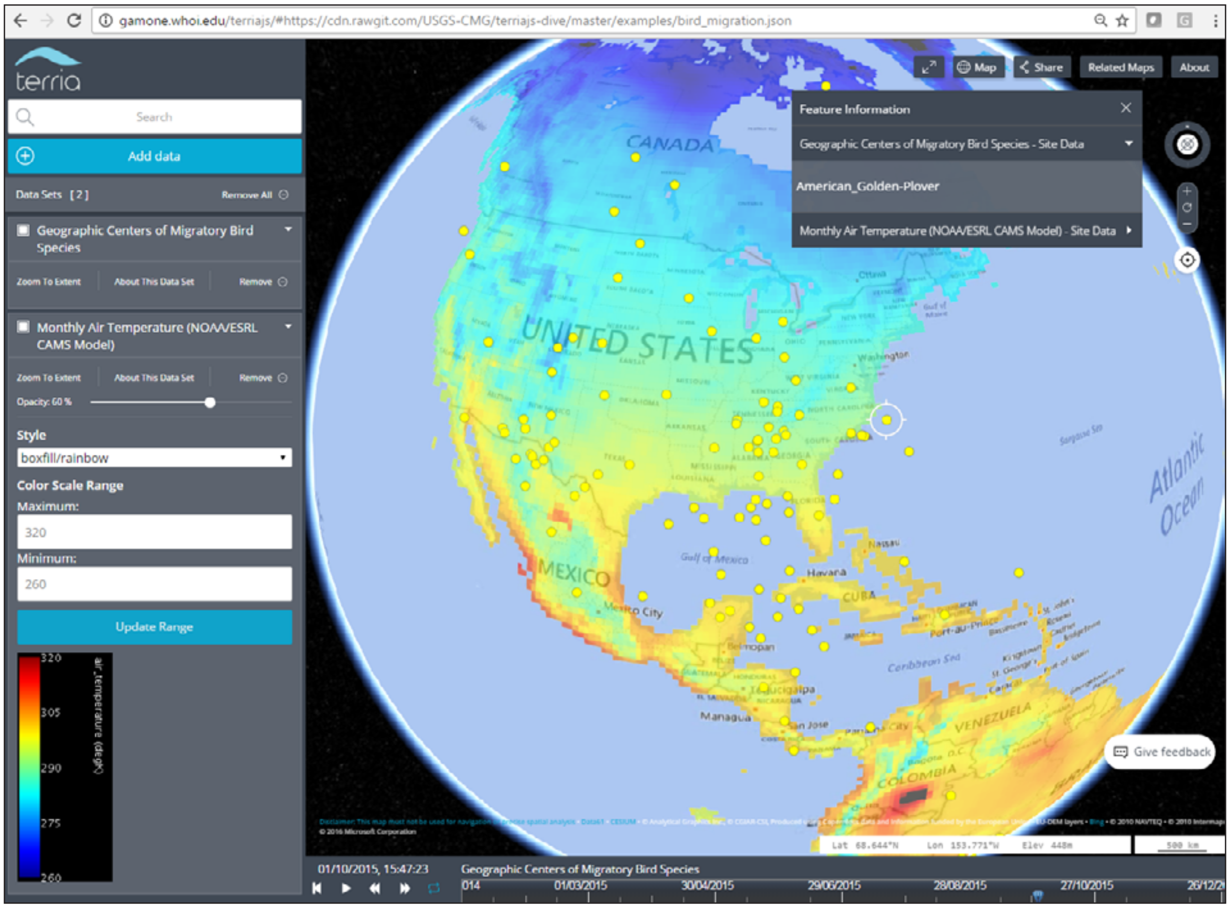
Minimum: 0

Disclaimer: This map must not be used for navigation or precise spatial analysis Data61 | Leaflet | Bing. Image courtesy of NASA, Earthstar Geographics SIO, © 2016 Microsoft Corporation, © 2010 NAVTEQ

Lat 43.616°N Lon 69.252°W Elev

19/10/2016, 14:56:19 u:v-group

Figure 11. Screen capture showing surface currents from the U.S. Integrated Ocean Observing System forecast model for New England, which uses a native triangular grid but is regridded at varying resolution by ncWMS2. Shown also are the user controls for selecting the layer, the elevation, and the color range. The ability to access ncWMS2 endpoints, as well as control the color range, elevation, and style, were all Community for Data Integration enhancements to TerriaJS. These enhancements have been merged into the master branch of TerriaJS.



gamone.whoi.edu/terriajs/#https://cdn.rawgit.com/USGS-CMG/terriajs-dive/master/examples/bird_migration.json

Data Sets [2]

Geographic Centers of Migratory Bird Species

Monthly Air Temperature (NOAA/ESRL CAMS Model)

Opacity: 60 %

Style: boxfill/rainbow

Color Scale Range

Maximum: 320

Minimum: 260

Feature Information

Geographic Centers of Migratory Bird Species - Site Data

American_Golden-Plover

Monthly Air Temperature (NOAA/ESRL CAMS Model) - Site Data

Lat 68.644°N Lon 153.771°W Elev 448m

01/10/2015, 15:47:23

Geographic Centers of Migratory Bird Species

Figure 12. Screen capture showing bird migration from a Cesium Markup Language (CZML) data source superimposed on monthly temperature data from web map service (via a National Oceanic and Atmospheric Administration THREDDS server). Shown also are the user controls added in this project for selecting the style and color range. Note that both the CZML file and the JavaScript Object Notation configuration for this demo are on GitHub at https://github.com/USGS-CMG/terriajs-dive and referenced directly in the URL of the portal. This means that users can create custom portals on their own, without any interaction from the provider of the TerriaJS endpoint, which just serves to deliver the TerriaJS code to the browser. TerriaJS runs completely within the browser.

Lessons Learned

Developing enhancements for TerriaJS was indeed challenging. USGS developer Jordan Walker found the documentation scarce, the code complex, and even building the software tricky due to numerous changing dependencies. Some of the challenges were caused by the fact that TerriaJS switched to a new user interface (UI) just after this CDI project was initiated. Most enhancements, therefore, ended up being developed by Kevin Ring, the lead developer for TerriaJS. As documentation and experience with the new UI grow, the situation with developing enhancements is expected to improve. A positive outcome, however, is that the new UI plays nicely on mobile devices, which is essential for uptake by the USGS community.

When enhancements were made to support the ncWMS and ncWMS2 services, the project team discovered that displaying information from unstructured grid (for example, triangular grid) models in TerriaJS was extremely slow and significantly slower than the Godiva3 ncWMS2 client. They discovered this because of a combination of factors: (1) slow delivery in general by the Java-based ncWMS2 service, (2) requesting tiles in projected coordinates rather than the native geographic coordinates, and (3) requesting many more tiles than Godiva3 for the same spatial extent. This discovery has triggered new, ongoing work to understand how to make delivery of information from unstructured grid models more efficient.

Although automatic deployment to the USGS infrastructure was successful, many of the external datasets were not displayed because of the USGS requirement of only accessing data from HTTPS. This project's use cases, designed to show integration of data from heterogeneous data services from the broad geospatial community, therefore, did not work because few external data providers supported HTTPS. It may be possible to work around this in the future by using a proxy. An additional weakness of the USGS TerriaJS endpoint is that it is only accessible via the internal USGS network. To overcome these shortcomings, the team set up a publicly accessible TerriaJS endpoint on a USGS/WHOI cooperative computer that doesn't have the HTTPS restrictions. This endpoint can be accessed at https://gamone.whoi.edu/terriajs.

The project team had hoped to bring in new USGS data from Dr. Patricia (Soupy) Dalyander to demonstrate integration of Docker in the TerriaJS platform, but security concerns with Docker prevented deployment of the THREDDS Data Server Docker container. The hope is that these security concerns are addressed by USGS in the future, as Docker makes deploying and maintaining services much easier, isolates server processes from each other, and makes services on specific machines more robust. The team would like to test deployment of the THREDDS Docker container on the USGS Cloud Hosting Solution in the near future.

Facilitating the USGS Scientific Data Management Foundation by Integrating the Process into Current Scientific Workflow Systems

Lead PIs: Colin B. Talbert, Drew A. Ignizio, Catherine Jarnevich, and Jeffrey T. Morisette

Increasing attention is being paid to the importance of proper scientific data management and implementing processes that ensure that products being released are properly documented. USGS policies have been established to properly document not only publications, but also the related data and software. This relatively recent expansion of documentation requirements for data and software may present a daunting challenge for many USGS scientists whose major focus is their physical science and who have less expertise in information science. As a proof of concept, this project has created a software solution that facilitates this process through a user-friendly, but comprehensive, interface embedded in an existing scientific workflow system used in the USGS for species distribution modeling. The software produced by this proposal has gone through initial testing, and the project team is currently in the process of using the application to document their first USGS data release. Lessons learned from this initial use case will be used to update and tune the current functionality. This functionality will ultimately be merged into the next released version of the Software for Assisted Habitat Modeling (SAHM). This project could provide an exemplar that the USGS can point to as it initiates new standards for producing repeatable science. The exemplar this team has created could also help ensure compliance pertaining to new requirements.

Accomplishments

The accomplishments for this project are described below.

- The explicit data management steps required to archive and document a workflow created with VisTrails/SAHM were identified and documented.
- The existing VisTrails/SAHM source code was extended to include tools to automate the steps in this process (fig. 13), such as
  - making clean copies of the relevant data files needed to reproduce the workflow, as well as the core outputs to create a self-contained archive bundle of the workflow's outputs;
  - using the MetadataWizard, now integrated directly into VisTrails/SAHM, to create FGDC-compliant metadata for the archive bundle (fig. 14); and
  - moving the completed archive bundle to ScienceBase and using the archive bundle's metadata to populate the item (fig. 15).
- The software and data management workflow were tested by a data manager at the USGS Fort Collins Science Center (FORT).
- The software is currently being used by FORT scientists to document a data release associated with a new publication. Contingent on the results of this test case, the team was planning to release the tool for wider, more comprehensive testing in the fall of 2016.

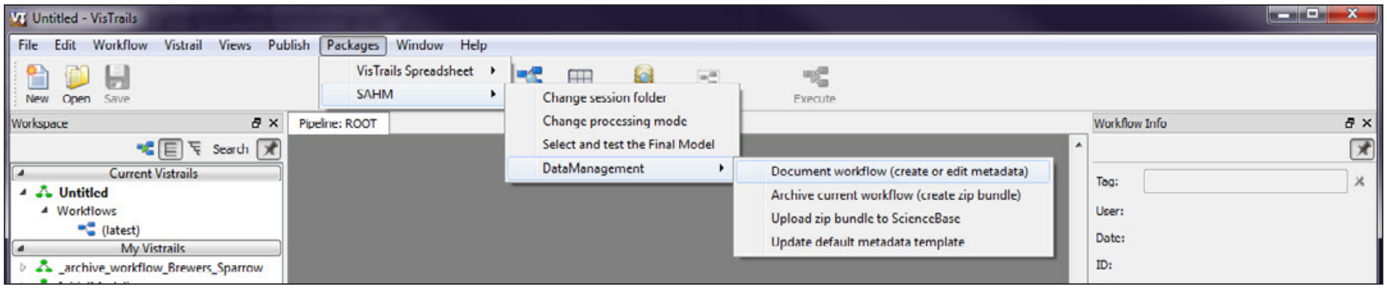
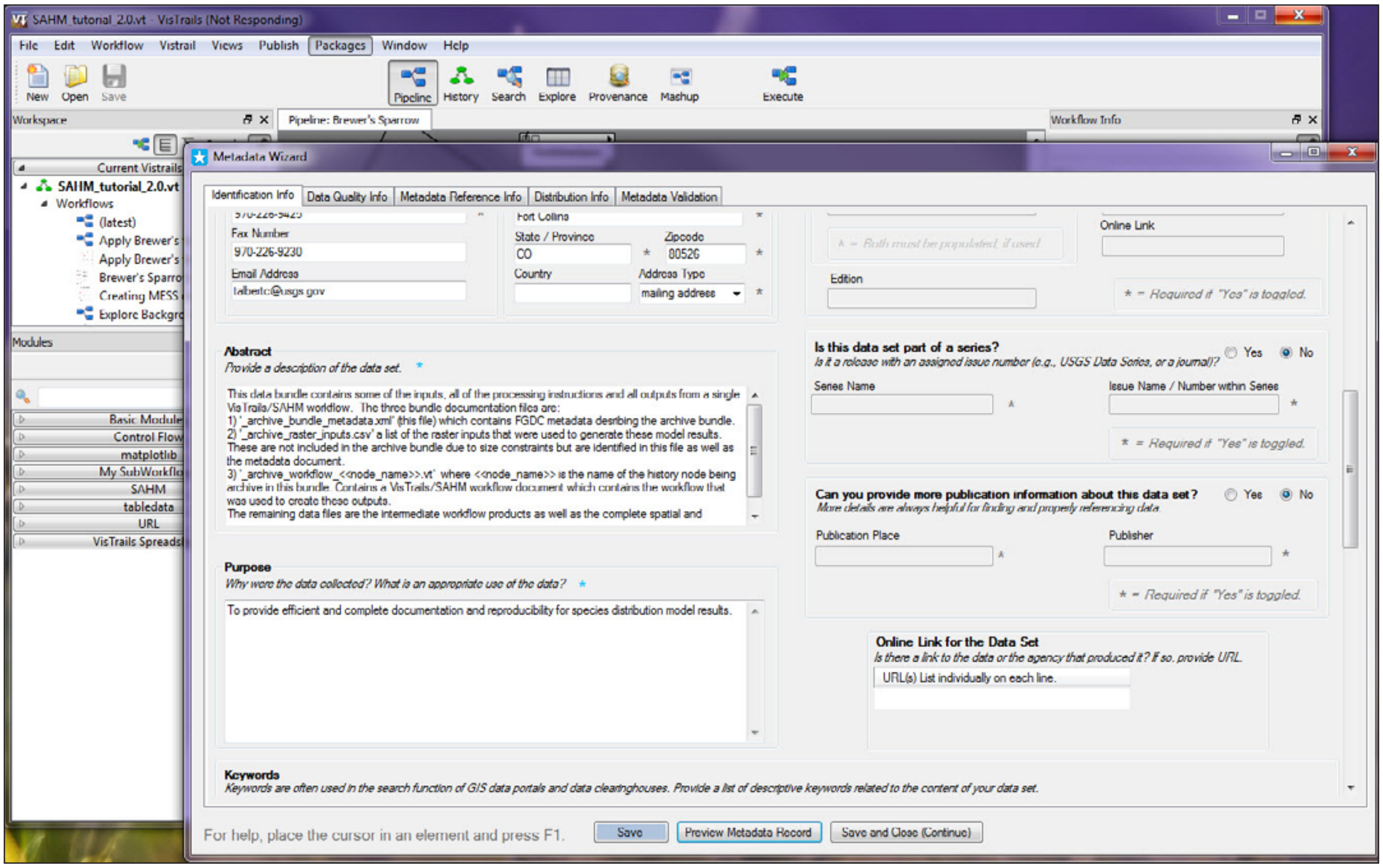

Untitled - VisTrails
File Edit Workflow Vistrail Views Publish Packages Window Help
New Open Save
VisTrails Spreadsheet
SAHM
Change session folder
Change processing mode
Select and test the Final Model
DataManagement
Document workflow (create or edit metadata)
Archive current workflow (create zip bundle)
Upload zip bundle to ScienceBase
Update default metadata template
Execute
Workspace
Search
Current Vistrails
Untitled
Workflows
(latest)
My Vistrails
_archive_workflow_Brewers_Sparrow
Pipeline: ROOT
Workflow Info
Tag:
User:
Date:
ID:

Figure 13. Dropdown menu of data management tasks integrated into VisTrails/Software for Assisted Habitat Modeling.



970-226-9425 *

Fax Number
970-226-9230

Email Address
talbertc@usgs.gov

Fort Collins *

State / Province
CO *

Zipcode
80526 *

Country

Address Type
mailing address *

Abstract
Provide a description of the data set. *

This data bundle contains some of the inputs, all of the processing instructions and all outputs from a single VisTrails/SAHM workflow. The three bundle documentation files are:
1) '_archive_bundle_metadata.xml' (this file) which contains FGDC metadata desribing the archive bundle.
2) '_archive_raster_inputs.csv' a list of the raster inputs that were used to generate these model results. These are not included in the archive bundle due to size constraints but are identified in this file as well as the metadata document.
3) '_archive_workflow_<<node_name>>.vt' where <<node_name>> is the name of the history node being archive in this bundle. Contains a VisTrails/SAHM workflow document which contains the workflow that was used to create these outputs.
The remaining data files are the intermediate workflow products as well as the complete spatial and

Purpose
Why were the data collected? What is an appropriate use of the data? *

To provide efficient and complete documentation and reproducibility for species distribution model results.

** = Both must be populated, if used.*

Online Link

Edition

** = Required if "Yes" is toggled.*

Is this data set part of a series? ○ Yes ● No
Is it a release with an assigned issue number (e.g., USGS Data Series, or a journal)?

Series Name *

Issue Name / Number within Series *

** = Required if "Yes" is toggled.*

Can you provide more publication information about this data set? ○ Yes ● No
More details are always helpful for finding and properly referencing data.

Publication Place *

Publisher *

** = Required if "Yes" is toggled.*

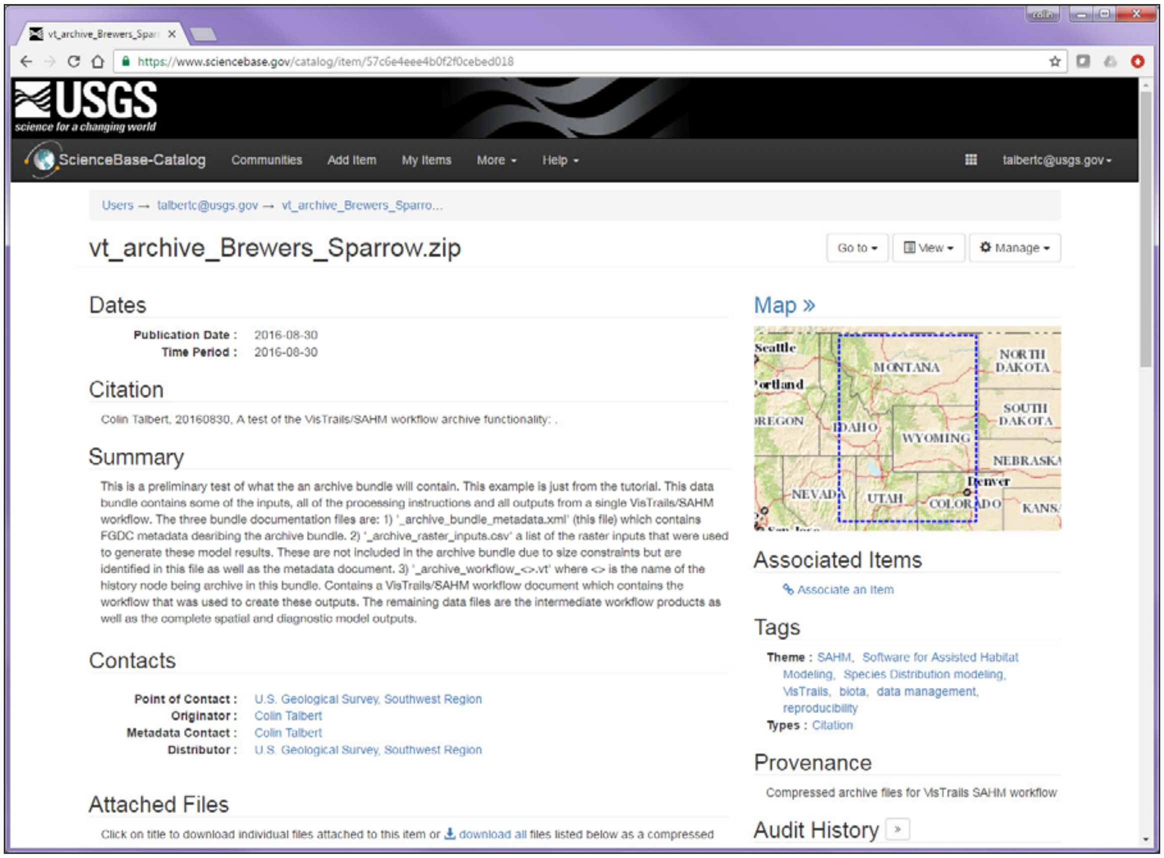
Online Link for the Data Set
Is there a link to the data or the agency that produced it? If so, provide URL.

URL(s) List individually on each line.

Keywords
Keywords are often used in the search function of GIS data portals and data clearinghouses. Provide a list of descriptive keywords related to the content of your data set.

For help, place the cursor in an element and press F1.

Figure 14. The Metadata Wizard launches with VisTrails to create Federal Geographic Data Committee metadata for an item.



vt_archive_Brewers_Sparrow.zip

Dates

Publication Date : 2016-08-30
Time Period : 2016-08-30

Citation

Colin Talbert, 20160830, A test of the VisTrails/SAHM workflow archive functionality: .

Summary

This is a preliminary test of what the an archive bundle will contain. This example is just from the tutorial. This data bundle contains some of the inputs, all of the processing instructions and all outputs from a single VisTrails/SAHM workflow. The three bundle documentation files are: 1) '_archive_bundle_metadata.xml' (this file) which contains FGDC metadata desribing the archive bundle. 2) '_archive_raster_inputs.csv' a list of the raster inputs that were used to generate these model results. These are not included in the archive bundle due to size constraints but are identified in this file as well as the metadata document. 3) '_archive_workflow_<>.vt' where <> is the name of the history node being archive in this bundle. Contains a VisTrails/SAHM workflow document which contains the workflow that was used to create these outputs. The remaining data files are the intermediate workflow products as well as the complete spatial and diagnostic model outputs.

Contacts

Point of Contact : U.S. Geological Survey, Southwest Region
Originator : Colin Talbert
Metadata Contact : Colin Talbert
Distributor : U.S. Geological Survey, Southwest Region

Attached Files

Click on title to download individual files attached to this item or download all files listed below as a compressed

Associated Items

Tags

Theme : SAHM, Software for Assisted Habitat Modeling, Species Distribution modeling, VisTrails, biota, data management, reproducibility
Types : Citation

Provenance

Compressed archive files for VisTrails SAHM workflow

Audit History

Figure 15. The archived workflow bundle can be posted to a user's items on ScienceBase.

Hunting Invasive Species with HTCondor: High Throughput Computing for Big Data and Next Generation Sequencing

Lead PIs: S. Grace McCalla, Michael Fienen, Richard Erickson, Randall Hunt, and Jon Amberg

Large amounts of data are being generated that require hours, days, or even weeks to analyze using traditional computing resources. Innovative solutions must be implemented to analyze the data in a reasonable timeframe. The program HTCondor (https://research.cs.wisc.edu/htcondor/) takes advantage of the processing capacity of individual desktop computers and dedicated computing resources as a single, unified pool. This unified pool of computing resources allows HTCondor to quickly process large amounts of data by breaking the data into smaller tasks distributed across many computers.

This project team implemented HTCondor at the USGS Upper Midwest Environmental Sciences Center (UMESC) to leverage existing computing capabilities for data processing and analysis. HTCondor can be used for a wide range of projects including processing DNA sequencing data (currently done as part of invasive species monitoring), validating new statistical models over a wide range of possible parameter combinations, and analyzing long-term vegetation and fish data from the Upper Mississippi River. The HTCondor pool is online and operational at UMESC. The USGS Wisconsin Water Science Center was able to connect to the pool at the USGS Oregon Water Science Center through "flocking" (fig. 16). This test identified cybersecurity and data transfer challenges that can be overcome in future work. Flocking with HTCondor requires communication among machines in various centers which, in turn, requires traffic to be allowed through the firewall of each center. When flocking takes place, many connections must be made, but all traffic is consolidated into a single port which is 9618. As a result, only port 9618 must be opened, and traffic can be limited to USGS computers. This minimizes the risk of allowing two-way traffic through firewalls from center to center. The technology underlying the test was shown to be successful and was an important step toward connecting and leveraging computing resources throughout the USGS.

Accomplishments

The accomplishments for this project are described below.

- HTCondor was installed and configured at UMESC.
- Configuration files and testing for flocking among USGS centers were created and performed.
- Documentation examples for using HTCondor to scale scientific processing with cluster computing were uploaded to USGS BitBucket at https://my.usgs.gov/bitbucket/projects/CDI/repos/hunting_invasive_species_with_htcondor/browse.
- Testing of Docker as a technology to exchange configuration of complex workflows and datasets throughout HTCondor pools was performed.

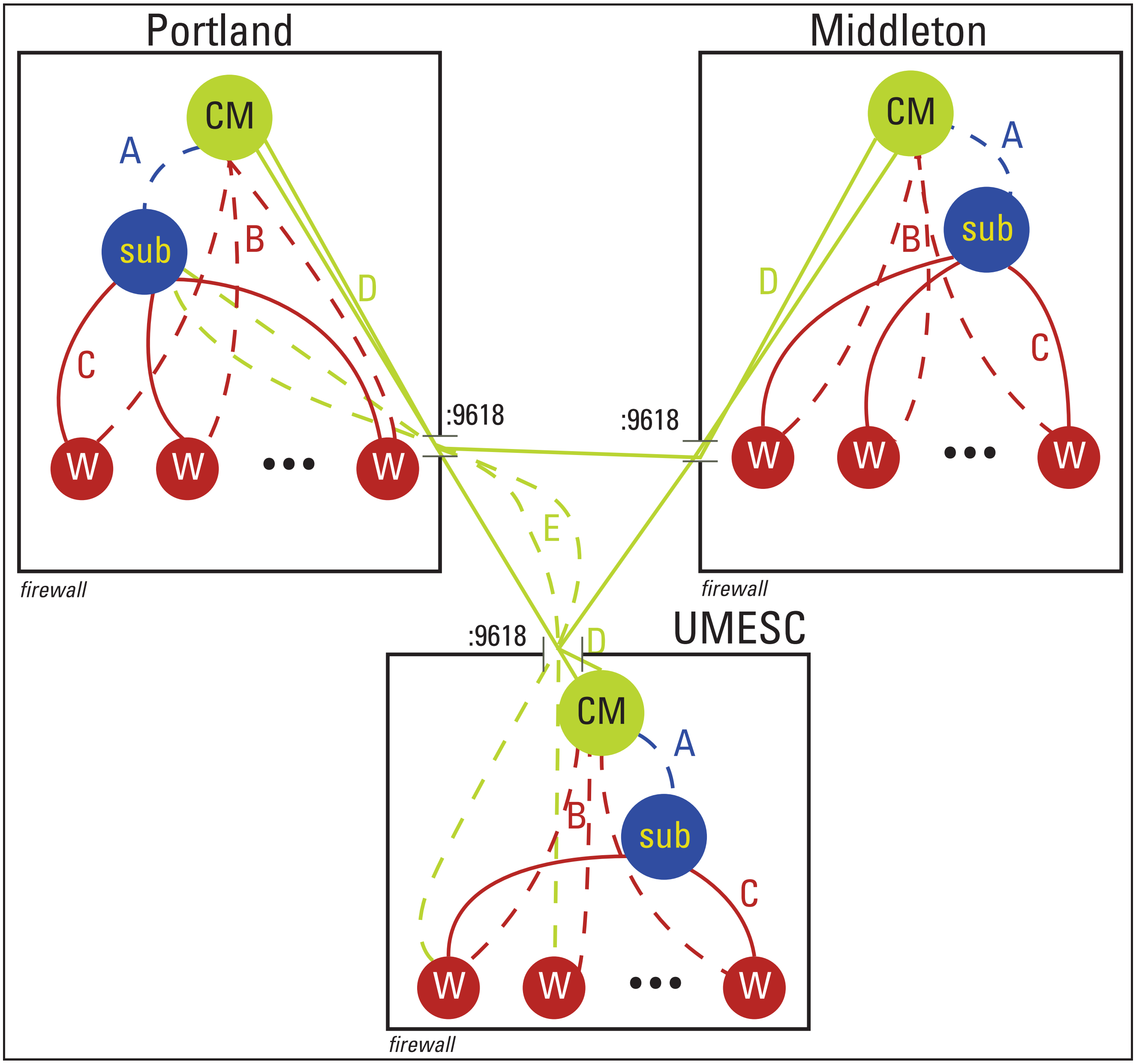


Figure 16. Schematic of flocking with HTCondor between U.S. Geological Survey centers. In the diagram, W = worker node, CM = central manager, and sub = submit node. The only port that needs to be opened is 9618 to enable flocking capabilities. A, B, C, D, and E are various connections.

Integration of National Soil and Wetland Datasets: A Toolkit for Reproducible Calculation and Quality Assessment of Imputed Wetland Soil Properties

Lead PIs: Eric T. Sundquist, Norman Bliss, Sharon Waltman, Rusty Griffin, and Lisamarie Windham-Myers

Wetland soils are vital to the Nation because of their role in sustaining water resources, supporting critical ecosystems, and sequestering significant concentrations of biologically produced carbon. The United States has the world's most detailed continent-scale digital datasets for soils and wetlands, yet scientists and land managers have long struggled with the challenge of integrating these datasets for applications in research and in resource assessment and management. The difficulties include spatial and temporal uncertainties, inconsistencies among data sources, and inherent structural complexities of the datasets. This project's objective was to develop and document a set of methods to impute wetland soil properties by integrating Soil Survey Geographic (SSURGO) data with the National Wetlands Inventory (NWI) and other data sources relevant to the extent and properties of wetlands.

The project methods build on the project team's current research and development of best practices for analysis and application of soil and wetland data. Documentation of the process is meant to assure complete transparency and reproducibility of imputed wetland soil properties, with a broad range of applications beyond the immediate interests of this project team.

Accomplishments

The project team combined the most recently available SSURGO dataset (mostly 1:24,000 map scale) with gap-filling from the generalized (1:250,000 map scale) Digital General Soil Map of the United States (STATSGO2) dataset. The combined SSURGO/STATSGO2 dataset is assembled as a map layer with associated tables of relational attributes for each soil map unit.

Using the combined SSURGO/STATSGO2 dataset, the team identified and extracted wetland-related attributes of soil map units, components, and component horizons for all soil map units within the conterminous United States. To facilitate imputation of wetland soil properties, the team represented these attributes in 10 spatially distinct map layers representing categorized occurrences of hydric soil components and soil flooding or ponding. The team also extracted 21 spatially distinct map layers from the NWI Wetlands Layer based on classification attributes related to soils and vegetation. Each map layer retains the full NWI alphanumeric classification code for each wetland polygon.

Preliminary spatial integration of the extracted SSURGO/STATSGO2 and NWI datasets has shown that individual soil map units with wetland-associated attributes often do not align spatially with established wetland polygons. The project team will continue to explore the spatial integration of these datasets, with emphasis on defining additional spatial relationships that will support imputation of wetland soil properties.

The project team created a geospatial data layer that supports imputation of SSURGO/STATSGO2 soil attributes for NWI wetlands based on a combination of spatial proximity and similarity of ecosystem and hydrographic settings. This "imputation layer" represents a combination of map layers representing 4-digit hydrologic unit codes and LANDFIRE/Nature Conservancy existing ecosystem valuation toolkit (EVT) and biophysical settings.

The team is conducting rigorous quality tests on the datasets described above, including area checksums, searches for null and illogical values, comparisons to independent datasets, and assessments of reproducibility reflected in similar recent data extractions for other investigations. These tests have revealed several unforeseen problems, which are now the primary focus of the team's ongoing work on this project.

Finally, the team has initiated inquiries concerning an appropriate domain repository for the project's datasets and accompanying documentation. They had originally planned for long-term storage and access of the data through the Oak Ridge National Laboratory Distributed Active Archive Center for Biogeochemical Dynamics (ORNL DAAC); however, it has recently become unclear whether the ORNL DAAC will continue archiving carbon-related datasets. The project team is waiting to learn more about this uncertainty before proceeding with data submission to the ORNL DAAC.

Integration of Phenological Forecast Maps for Assessment of Biodiversity: An Enterprise Workflow

Lead PIs: Jake F. Weltzin, Alyssa Rosemartin, R. Lee Marsh, R. Sky Bristol, Tim Kern, and Theresa M. Crimmins,

Recent open data policies of the Office of Science and Technology Policy (OSTP) and Office of Management and Budget (OMB), which were fully enforceable on October 1, 2016, require that federally funded information products (publications, etc.) be made freely available to the public and that the underlying data on which the conclusions are based must be released. A key and relevant aspect of these policies is that data collected by USGS programs must be shared with the public and that these data are subject to the review requirements of Fundamental Science Practices (FSP). These new policies add a substantial burden to USGS scientists and science centers; however, the upside of working towards compliance with top-down policies is improved discovery, accessibility (including machine readability), integration, and use of data for novel applications in support of the broader USGS mission. The purpose of this research was to exercise these new policies, as they relate to production of real-time and short-term forecasts of gridded biodiversity data, as a model for similar production and delivery of data products from other USGS projects and programs.

The objectives of this project were as follows.

1. Establish and document a generalized workflow for making USGS biodiversity data and data products available for use by other programs within the USGS as well as by external partners and the public.
2. Develop technical documentation and open-source code, enabling others to reuse infrastructure for data processing, validation, delivery, and reporting on data usage.
3. Document this generalized workflow and lessons learned to significantly advance the delivery and translation of critical ecological information for a variety of stakeholders within the Department of the Interior (DOI) and beyond in a USGS open-file report (OFR).
4. Produce and release dynamic gridded maps, historical, real-time, and short-term forecasted predictions of leaf-out and flowering for several plant species and underlying accumulated temperature data products through a case study serving as a test bed for developing a generalized workflow for real-time delivery of dynamic biodiversity data products.

Accomplishments

The accomplishments for this project are described below.

- The project team created, documented, and published two suites of national-scale gridded maps: Extended Spring Indices and Accumulated Growing Degree Days. The data, metadata, and documentation have undergone FSP review and were released in early FY 2017 (Crimmins and others, 2017).
- The team implemented OGC-compliant web services serving the gridded data and initiated collaboration with the National Biogeographic Map (NBM) under development and FSP review by Sky Bristol. The Extended Spring Index layers have been brought into the NBM as a “Bioscape,” and the Accumulated Growing Degree Day layers will also be brought into the tool soon. The project team continues to work with the NBM team to develop data summaries, called Bioscape Analysis Packages, or BAPs, that will enable NBM users to explore patterns in the Spring Indices within National Park Service and National Wildlife Refuge System units (Parks and Refuges) and also to look at daily growing degree day accumulations within a pixel over the course of a year. These data layers and BAPs will be part of the NBM when it goes live in early 2017.
- The python scripts and workflow associated with producing these suites of gridded data layers were checked into the USGS BitBucket code repository (https://my.usgs.gov/bitbucket/projects/CDI/repos/phenology-maps-workflow/browse).
- The team updated the USA National Phenology Network (NPN) ScienceBase entry to include information on the two suites of gridded data products (https://www.sciencebase.gov/catalog/item/52fd3728e4b0f010068e97ce).
- The team developed an API to share their data layers with the University of Arizona Library Spatial Data Explorer (https://geo.library.arizona.edu/).
- The team developed upgrades to the online USA-NPN Visualization tool to show the gridded layers as maps, which can be viewed independently or in conjunction with in-situ plant or animal phenology observation data (https://www.usanpn.org/data/visualizations).
- Map images (.png, .gif, .pdf) (figure 17) or Web Coverage Service (WCS) and WMS raster data files (GeoTiff, ArcGrid, NetCDF) are freely available and can be downloaded using the USA-NPN Geoserver Request Builder page (https://www.usanpn.org/geoserver-request-builder).
- Web services are available via the USA-NPN Geoserver (http://geoserver.usanpn.org/geoserver/wms?request=GetCapabilities).
- Additional interpretive material is available at https://www.usanpn.org/data/phenology_maps.
- The project team is preparing a USGS OFR that will document a generalized workflow and lessons learned to significantly advance the delivery and translation of critical ecological information for a variety of stakeholders within the Department of the Interior and beyond. This OFR will include documentation of how leaders of a USGS project would

review and validate their data products against each component of the data lifecycle and how they might prioritize efforts to reach compliance based on the OSTP and OMB open data policies. The audience for this report would be USGS project leaders seeking to publish and share their data and data products.

- The incipient publication of the USGS Survey Manual chapter 502.8 provides broad guidance for the review and approval of scientific data for release but insufficient documentation to readily identify a workflow for review and approval of dynamic data released via APIs, or web services. The team is continuing to collaborate with members of the Office of Science Quality and Integrity, members of the Core Science Analytics, Synthesis and Library program, and the Fundamental Science Practices Advisory Committee on such a workflow.
- A Technical Information Sheet describing the gridded products was completed, reviewed, and published (USA National Phenology Network, 2016).

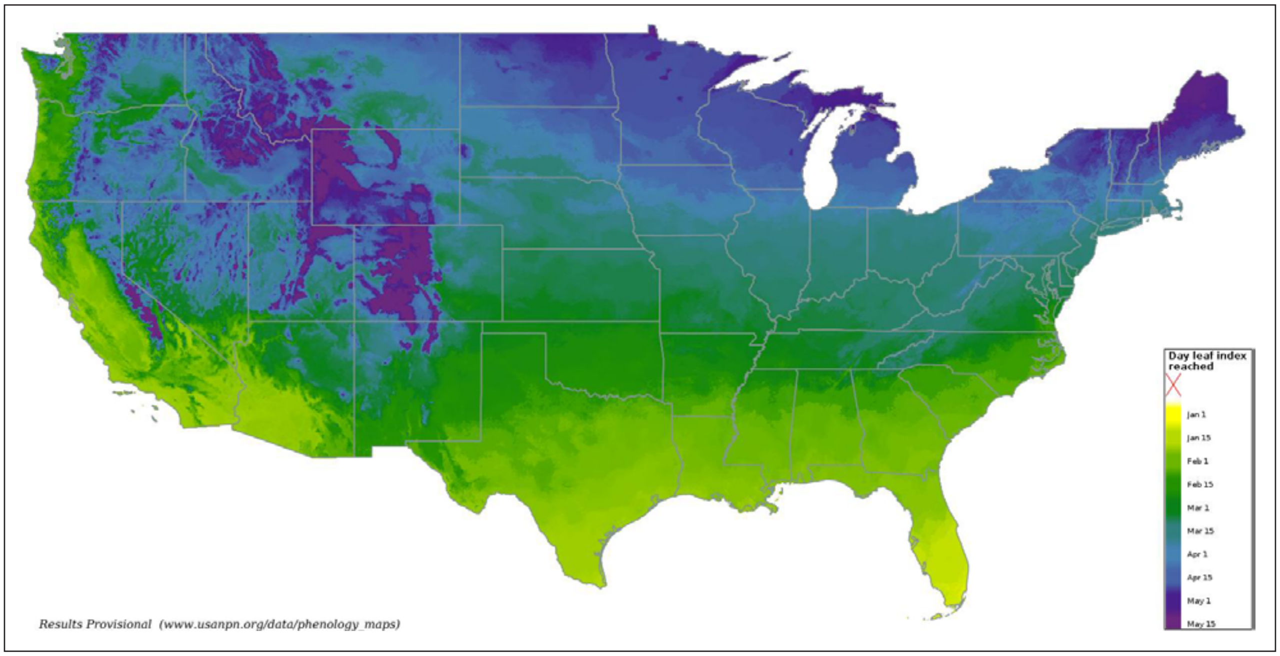


Figure 17. 2016 Spring Index First Leaf threshold dates as of October 1, 2016 (results provisional).

National Stream Summarization: Standardizing Stream-Landscape Summaries

Lead PIs: Daniel J. Wieferich, Jeff Falgout, Dana M. Infante, Scott Leibowitz, Marc Weber, and Brad Williams

As research and management of natural resources shift from local to regional and national scales, the need for information about aquatic systems to be summarized to multiple scales is becoming more apparent. Recently, four federally funded national stream assessment efforts (USGS Aquatic GAP, USGS National Water-Quality Assessment Program, U.S. Environmental Protection Agency [EPA] StreamCat, and National Fish Habitat Partnership) identified and summarized landscape information into two hydrologically and ecologically significant scales of local and network catchments for the National Hydrography Dataset Plus (NHDPlus). These efforts have revealed a significant percentage of assessment funds being directed to the collection and processing of data instead of for the assessments themselves. Additionally, although similar data are being summarized across these efforts, each is creating its own implementation. This duplication of effort is inefficient and may be producing inconsistent results.

To address these issues, the USGS, EPA, and Michigan State University participants have used CDI funds to support progress towards collaborative efforts with the end goal of developing a common workflow (for example, code) to accurately and efficiently summarize landscape information into local and network catchments of the NHDPlus.

Accomplishments

Current accomplishments and continued efforts are described below.

- The project team documented comparisons of existing workflows and resulting datasets to help identify commonalities and differences of current efforts that need to be addressed as a standardized workflow is developed. Detailed comparisons can be found at https://my.usgs.gov/confluence/x/2IoJI.
- An interagency agreement was established to outline and facilitate a 1-year commitment between the EPA and USGS where EPA staff will assist in the implementation, development, and refinement of a standardized software to process stream summarization (FY 2017 Planned Accomplishment). The interagency agreement will also facilitate a proposal for a single, common dataset repository and metadata documentation of landscape information summarized to the NHDPlusV2 (FY 2017 Planned Accomplishment).
- A face-to-face meeting took place during August 3–5, 2016, to compare current stream summarization efforts, understand needs of potential user groups across the Federal government, and identify a path forward. Details are documented at https://my.usgs.gov/confluence/x/2IoJI.
- A confluence page was developed to help build a community of practice and to help disseminate and coordinate current and future efforts related to the CDI project. The page can be accessed at https://my.usgs.gov/confluence/x/lIcJI.

Summary

The grassroots nature of the Community for Data Integration (CDI) has once again enabled the community to accomplish a tremendous amount of hard work in fiscal year (FY) 2016. Through the monthly forums, workshops, working groups, projects, and most importantly by constant surveying of the community's needs, the CDI has provided valuable content that keeps current members engaged and attracts new members.

The CDI strives to be a resource for any person or group that is looking for a community framework to support and enhance their data and science activities. The community answers requests to help support U.S. Geological Survey (USGS) events such as the Mapping Innovation Workshop Series, hosts joint events with other established groups such as during the joint USGS and Federation of Earth Science Information Partners (ESIP) Hackathon, and continues to fund diverse projects that add value to USGS data assets and tools. As the CDI has matured since its inception in 2009, it has grown to have a presence in the wide spectrum of disciplines and geographic locations represented in USGS and its partners.

Continuing to grow awareness of its activities and products and to reach every discipline and location are ongoing goals of the CDI. Through participation in ESIP, Research Data Alliance, and professional society meetings, the CDI finds points of connection and collaboration with other groups working on similar data challenges. These communications benefit the CDI membership by bringing new opportunities to light and finding places where efforts can be leveraged and shared. In this way, the trainings, resources, and networking provided by the CDI can truly affect the advancement of data and science by the USGS.

References Cited

Burns, Sylvia, 2016, Department of the Interior mobile applications privacy policy: U.S. Department of the Interior, OCIO Directive 2016–003, accessed December 27, 2016, at https://www.doi.gov/sites/doi.gov/files/uploads/ocio_directive_2016-003_doi_mobile_applications_privacy_policy.pdf.

Chase, K.J., Bock, A.R., and Sando, Roy, 2017, Sharing our data—An overview of current (2016) USGS policies and practices for publishing data on ScienceBase and an example interactive mapping application: U.S. Geological Survey Open-File Report 2016–1202, 10 p., accessed March 21, 2017, at https://doi.org/10.3133/ofr20161202.

Chase, K.J., Haj, A.E., Regan, R.S., and Viger, R.J., 2016a, Potential effects of climate change on streamflow for seven watersheds in eastern and central Montana: Journal of Hydrology—Regional Studies, v. 7, p. 69–81, accessed August 5, 2016, at http://dx.doi.org/10.1016/j.ejrh.2016.06.001.

Chase, K.J., Haj, A.E., Regan, R.S., and Viger, R.J., 2016b, Potential effects of climate change on streamflow in eastern and central Montana (2013–2014 analyses)—PRMS input and output data: U.S. Geological Survey data release, accessed November 16, 2016, at http://dx.doi.org/doi:10.5066/F7P26W5S.

Cheng, Joe, and Xie, Y., 2016, Leaflet—Create interactive web maps with the Javascript 'Leaflet' Library: R package version 1.0.1, accessed November 16, 2016, at https://CRAN.R-project.org/package=leaflet.

Community for Data Integration, 2016, CDI projects fiscal year 2016: U.S. Geological Survey, accessed November 17, 2016, at https://www.sciencebase.gov/catalog/item/56d86f51e4b015c306f6cf8b.

Crimmins, T.M., Marsh, R.L., Switzer, J.R., Crimmins, M.A., Gerst, K.L., Rosemartin, A.H., and Weltzin, J.F., 2017, USA National Phenology Network gridded products documentation: U.S. Geological Survey Open-File Report 2017–1003, 27 p., accessed March 21, 2017, at https://doi.org/10.3133/ofr20171003.

Earthquake Hazards Program, 2016, M 2.9—13km NE of Quito, Ecuador: U.S. Geological Survey, accessed November 3, 2016, at http://earthquake.usgs.gov/earthquakes/eventpage/us10006tlr#executive.

Faundeen, J.L., Burley, T.E., Carlino, J.A., Govoni, D.L., Henkel, H.S., Holl, S.L., Hutchison, V.B., Martín, Elizabeth, Montgomery, E.T., Ladino, C.C., Tessler, Steven, and Zolly, L.S., 2013, The United States Geological Survey Science Data Lifecycle Model: U.S. Geological Survey Open-File Report 2013–1265, 4 p., accessed February 3, 2017, at http://dx.doi.org/10.3133/ofr20131265.

Kern, Tim, 2015, Mobile application release checklist: U.S. Geological Survey, accessed October 27, 2016, at https://my.usgs.gov/confluence/x/JInQHw.

Langseth, M.L., Chang, M.Y., Carlino, Jennifer, Bellmore, J.R., Birch, D.D., Bradley, Joshua, Bristol, R.S., Buscombe, D.D., Duda, J.J., Everette, A.L., Graves, T.A., Greenwood, M.M., Govoni, H.S., Henkel, H.S., Hutchison, V.B., Jones, B.K., Kern, Tim, Lacey, Jennifer, Lamb, R.M., Lightsom, F.L., Long, J.L., Saleh, R.A., Smith, S.W., Soulard, C.E., Viger, R.J., Warrick, J.A., Wesenberg, K.E., Wieferich, D.J., and Winslow, L.A., 2016, Community for Data Integration 2015 annual report: U.S. Geological Survey Open-File Report 2016–1165, 57 p., accessed October 27, 2016, at http://dx.doi.org/10.3133/ofr20161165.

Lightsom, F., Allwardt, A., Schweitzer, P., and Zolly, L., 2015, The controlled vocabulary manifesto: U.S. Geological Survey, accessed October 25, 2016, at https://my.usgs.gov/confluence/display/cdi/SWWG+2014+Proposal+Project?preview=/477659162/541727879/Controlled_Vocabulary_Manifesto_20151029_for_Confluence.pdf

Schweitzer, P.N., 1995, MP—A compiler for formal metadata, edition 2.9.34: U.S. Geological Survey software, accessed October 21, 2016, at http://geology.usgs.gov/tools/metadata/.

TerriaJS, 2015, TerriaJS: GitHub repository, accessed November 16, 2016, at https://github.com/TerriaJS/terriajs.

USA National Phenology Network, 2016, Daily accumulated growing degree day and Spring Index maps: USA-NPN Technical Information Sheet, accessed March 22, 2017, at https://usanpn.org/files/shared/files/USA-NPN-AGDD-and-SiX.pdf.

U.S. Geological Survey, 2015, U.S. Geological Survey Community for Data Integration (CDI) science support framework (SSF): U.S. Geological Survey, 3 p., accessed November 18, 2016, at http://www.usgs.gov/cdi/cdi-ssf/cdi-ssf-components.pdf.

U.S. Geological Survey, 2016, Public access to results of federally funded research at the U.S. Geological Survey—Scholarly publications and digital data: U.S. Geological Survey, accessed October 24, 2016, at https://www2.usgs.gov/quality_integrity/open_access/downloads/USGS-PublicAccessPlan-APPROVED-v1.03.pdf.

Winslow, Luke, Chamberlain, S., Appling, A., and Read, J., 2016, sbtools—Tools for interfacing R with ScienceBase data services: U.S. Geological Survey, accessed November 16, 2016, at https://github.com/USGS-R/sbtools. [For use with R package version 1.0.2.]

Wood, N.J., and Jones, J.M., 2017, Tsunami travel time maps for Del Norte and Humboldt Counties, CA, reference year 2010: U.S. Geological Survey data release, accessed March 31, 2017, at https://doi.org/10.5066/F7CC0XWN.

Langseth and others—**Community for Data Integration 2016 Annual Report**— Open-File Report2 2017–1053

ISSN 2331-1258 (online)
https://doi.org/10.3133/ofr20171053