



Community for Data Integration 2014 Annual Report

By Madison L. Langseth, Michelle Y. Chang, Jennifer Carlino, Daniella D. Birch, Joshua Bradley, R. Sky Bristol, Craig Conzelmann, Robert H. Diehl, Paul Earle, Laura E. Ellison, Anthony L. Everette, Pam Fuller, Janice M. Gordon, David L. Govoni, Michelle R. Guy, Heather S. Henkel, Vivian B. Hutchison, Tim Kern, Frances L. Lightsom, Joseph W. Long, Ryan Longhenry, Todd M. Preston, Stan Smith, Roland J. Viger, Katherine Wesenberg, and Eric Wood

Open-File Report 2015–1184

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

U.S. Department of the Interior
SALLY JEWELL, Secretary

U.S. Geological Survey
Suzette M. Kimball, Acting Director

U.S. Geological Survey, Reston, Virginia: 2015

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://www.usgs.gov/pubprod>

To order this and other USGS information products, visit <http://store.usgs.gov>

Suggested citation:

Langseth, M.L., Chang, M.Y., Carlino, Jennifer, Birch, D.D., Bradley, Joshua, Bristol, R.S., Conzelmann, Craig, Diehl, R.H., Earle, Paul, Ellison, L.E., Everette, A.L., Fuller, Pam, Gordon, J.M., Govoni, D.L., Guy, M.R., Henkel, H.S., Hutchison, V.B., Kern, Tim, Lightsom, F.L., Long, J.W., Longhenry, Ryan, Preston, T.M., Smith, Stan, Viger, R.J., Wesenberg, Katherine, and Wood, Eric, 2015, Community for Data Integration 2014 annual report: U.S. Geological Survey Open-File Report 2015-1184, 40 p., <http://dx.doi.org/10.3133/ofr20151184>.

ISSN 2331-1258 (online)

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

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

Contents

Abstract	1
Community for Data Integration Science Support Framework	2
Monthly Forums and Annual Workshop/Webinar Series	3
Community for Data Integration Monthly Forums	3
2014 Community for Data Integration Workshop and Webinar	5
Working Groups and Focus Groups	6
Citizen Science Working Group	6
Citizen Science Working Group Accomplishments	7
Connected Devices Working Group	7
Connected Devices Working Group Accomplishments	7
Connected Devices Working Group Meetings and Presentations	7
Data Management Working Group	8
Data Management Working Group Accomplishments	8
Data Management Web site	8
Science Data Lifecycle Model	9
USGS Science Data Exit Survey Form	9
Data Policy Writing Team	9
Data Release Team	10
Digital Object Identifiers (DOIs) Team	10
Data Management Working Group Meetings and Presentations	10
Data Management Working Group Feedback from Stakeholders	10
Semantic Web Working Group	11
Semantic Web Working Group Accomplishments	11
Semantic Web Working Group Meetings and Presentations	12
Technology Stack Working Group	13
Technology Stack Working Group Accomplishments	13
Annual Community for Data Integration Request for Proposals	13
Phase I—Statements of Interest	14
Phase II—Full Proposals	15
Review Panel Process	15
Bureau-Wide Applications	15
Recommendations	16
Community for Data Integration Projects	17
Adopt-a-Pixel—Data Infrastructure	17
Deliverables	17
Characterization of Earthquake Damage and Effects Using Social-Media Data	18
Deliverables	19
Geographic Searches for USGS Publications	23
Deliverables	24
NASWeb Application Programming Interface—Web Services Access to the Nonindigenous Aquatic Species Database	25
Deliverables	26
North American Bat Data Integration	27
Deliverables	27
Online Merging and Gridding of Topographic and Bathymetric Data Sources	28
Deliverables	29

Portable ISO 19115 Open Source Developer's Toolkit	30
Deliverables.....	31
Summarization of National NEXt Generation RADar Data for Use in Biological Applications	33
Deliverables.....	35
Use of Controlled Vocabularies in USGS Information Applications.....	36
Deliverables.....	37
Summary	37
References Cited	38

Figures

1. Community for Data Integration Science Support Framework.....	2
2. U.S. Geological Survey (USGS) data management Web site home page.....	9
3. A diagram showing the three controlled vocabulary use cases and their relationships to each other.....	11
4. The Adopt-a-Pixel Web interface shows uploaded imagery from a single point facing north, south, east, west, up, and down to give a clear view of landscape features in all directions	18
5. This figure shows the earthquake significance summary report that is automatically derived and emailed 5 and 10 minutes after the event detection	20
6. This figure shows a screenshot of the Kibana interactive visualization and data analysis tool which includes charts showing the event detection time based on a spike in the number of tweets over time (top), the language of the tweet (left center), keywords used that indicate the earthquake's shaking intensity and effects (center), and the type of tweet (right center). The tool's interface also includes a pinned tweet map, which shows the tweet locations and the density of tweets at those locations (bottom)	21
7. This figure shows the Kibana interactive visualization and data analysis tool workflow. The event photos, which support shaking and damage estimates, shown in the figure were tweeted from the locations shown on the map. These photos appear when the tweet map location is clicked	22
8. Heatmap produced from ArcGIS service shows the distribution and density reporting tweets	23
9. Screenshot of ScienceBase Footprint Studio displaying the footprint for the Ocean Biogeographic Information System (OBIS) – United States of America Dataset Collection	24
10. Screenshot of the PubsWarehouse advanced search function, which now allows geospatial footprint queries.....	25
11. Comparison of the NASWeb Web-accessible specimen record (left) with the textual version that is returned from the NASWeb application programming interface (right)	26
12. A screen capture of the spread of white-nose syndrome in North America	28
13. Framework for online, on-demand tool to create coastal digital elevation models. Users define an area of interest using the example Web interface (top panel). Based on a query of elevation resources, relevant datasets in the selected region is identified (middle panel) where, in this example, bathymetric data are shown in red and locations of topographic data in black. Finally, the geoprocessing service merges, interpolates, and smooths the data into a digital elevation model using the user-defined requirements (bottom panel)	30
14. This figure briefly describes each component of the Portable ISO 19115 Open Source Developer's Toolkit and illustrates some possible workflows for generating ISO 19115 metadata using the toolkit	33
15. Examples of biological targets detected by NEXRAD	34
16. Steps in the process of enabling the biological application of weather radar from data collection and archiving, to postprocessing, to end user	35
17. Summarized NEXRAD data from 21 radar stations showing the strength of radar echoes associated with bird migration (circular patterns) in the presence of three strong tropical storms that made landfall during the 2005 hurricane seasons, Katrina (upper left), Rita (upper right), and Wilma (bottom)	36

Tables

1.	List of 2014 Community for Data monthly forums	4
2.	Community for Data Integration 2014 working groups and focus groups.....	6
3.	Number of proposals submitted for the Community for Data Integration fiscal year 2014 request for proposals by Science Support Framework category.....	14
4.	Overview of the Community for Data Integration request for proposals projects funded in fiscal year 2014 (in alphabetical order).....	16
5.	Analyses of the integrated social media based event detections and seismically derived events.....	23

Abbreviations

ADIWG	Alaska Data Integration working group
API	Application programming interface
BPD	USGS Bat Population Data
BCO-DMO	Biological and Chemical Oceanography Data Management Office
BISON	Biodiversity Information Serving Our Nation
CDI	Community for Data Integration
CDWG	Connected Devices Working Group
CIDA	Center for Integrated Data Analytics
CLI	Command-line interface
CMGP	Coastal and Marine Geology Program
CSAS&L	Core Science Analytics, Synthesis & Libraries
CSDGM	Content Standard for Digital Geospatial Metadata
CSV	Comma separated value
CSWG	Citizen Science Working Group
DEM	Digital elevation model
DMWG	Data Management Working Group
DOIs	Digital object identifiers
DOI	U.S. Department of the Interior
ESIP	Federation of Earth Science Information Partners
ESRI	Environmental Systems Research Institute
FGDC	Federal Geographic Data Committee
FORT	Fort Collins Science Center
FSP	Fundamental Science Practices
GIS	Geographic information systems
GML	Geography Markup Language
ISO	International Organization for Standardization
IT	Information technology
JSON	JavaScript Object Notation
NAS	Nonindigenous aquatic species
NCDC	National Climatic Data Center

NEXRAD	NEXt Generation RADar
NGO	Nongovernmental organization
NOAA	National Oceanic and Atmospheric Administration
NSF	U.S. National Science Foundation
OGC	Open Geospatial Consortium
OMB	Office of Management and Budget
OSTP	Office of Science and Technology Policy
RFP	Request for proposals
RPI TWC	Rensselaer Polytechnic Institute Tetherless World Constellation
SDLC	Science data life cycle
SOI	Statement of interest
SSF	Science Support Framework
SWWG	Semantic Web Working Group
TSWG	Tech Stack Working Group
UNESCO	United Nations Educational, Scientific and Cultural Organization
USGS	U.S. Geological Survey
WNS	White-nose syndrome

Community for Data Integration 2014 Annual Report

By Madison L. Langseth, Michelle Y. Chang, Jennifer Carlino, Daniella D. Birch, Joshua Bradley, R. Sky Bristol, Craig Conzelmann, Robert H. Diehl, Paul Earle, Laura E. Ellison, Anthony L. Everette, Pam Fuller, Janice M. Gordon, David L. Govoni, Michelle R. Guy, Heather S. Henkel, Vivian B. Hutchison, Tim Kern, Frances L. Lightsom, Joseph W. Long, Ryan Longhenry, Todd M. Preston, Stan Smith, Roland J. Viger, Katherine Wesenberg, and Eric Wood

Abstract

The U.S. Geological Survey (USGS) researches Earth science to help address complex issues affecting society and the environment. In 2006, the USGS held the first Scientific Information Management Workshop to bring together staff from across the organization to discuss the data and information management issues affecting the integration and delivery of Earth science research and investigate the use of “communities of practice” as mechanisms to share expertise about these issues. Out of this effort emerged the Council for Data Integration, which was conceived as an official organizational function that would help guide data integration activities and formalize communities of practice into working groups; however, by 2009 it became evident that many members of the Council for Data Integration had an interest in developing data integration solutions and sharing expertise in a less formal, grassroots manner, which transformed the Council into a Community for Data Integration (CDI). As of 2014, the CDI represents a dynamic community of practice focused on advancing science data and information management and integration capabilities across the 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). The purpose of the CDI is to do the following:

- advance understanding of Earth systems through enhanced use of data and information including associated tools and techniques,
- provide a forum for people doing work with data integration to come together to share ideas and learn new skills and techniques, and
- grow overall USGS capabilities with data and information by increasing visibility of the work of many people throughout the USGS and the CDI community.

To achieve these goals, the CDI operates within four applied areas: monthly forums, annual workshop/webinar series, working groups, and projects. The monthly forums, also known as the Opportunity/Challenge of the Month, 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. Since 2010, the CDI has also sponsored annual workshops/webinar series to encourage the exchange of ideas, sharing of activities, presentations of current projects, and networking among members. Stemming from common interests, the working groups are focused 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. The growing support for the activities of the working groups led to the CDI's first formal request for proposals (RFP) process in 2013 to fund projects that produced tangible products. As of 2014, the CDI continues to hold an annual RFP that creates data management tools and practices, collaboration tools, and training in support of data integration and delivery.

Community for Data Integration Science Support Framework

In order to provide an overarching context and vision for CDI goals and activities, the CDI Coordinators, consisting of working group leads and facilitators, developed the Science Support Framework (SSF) in 2012 (U.S. Geological Survey, 2015a). The SSF categorizes the activities and processes through which research data flow and upon which the CDI operates. It is these categories that provide the operational foundation and conceptual architecture that illustrates how CDI activities contribute to Bureau-level data integration efforts.

The vertical elements in the SSF (fig. 1) represent the “how” of the CDI—processes; implementation of standards and best practices; and interactions among people, data, and technology necessary to achieve data integration. The activities of monitoring, assessing, and researching flow through the science data life cycle (Faundeen and others, 2013) processes with the aid of applications, Web services, and semantics (that is, common frameworks and ontologies for sharing data across applications, communities, enterprises, and so forth). The assets are transformed into information products that increase knowledge and understanding of the Earth's physical and biological systems.

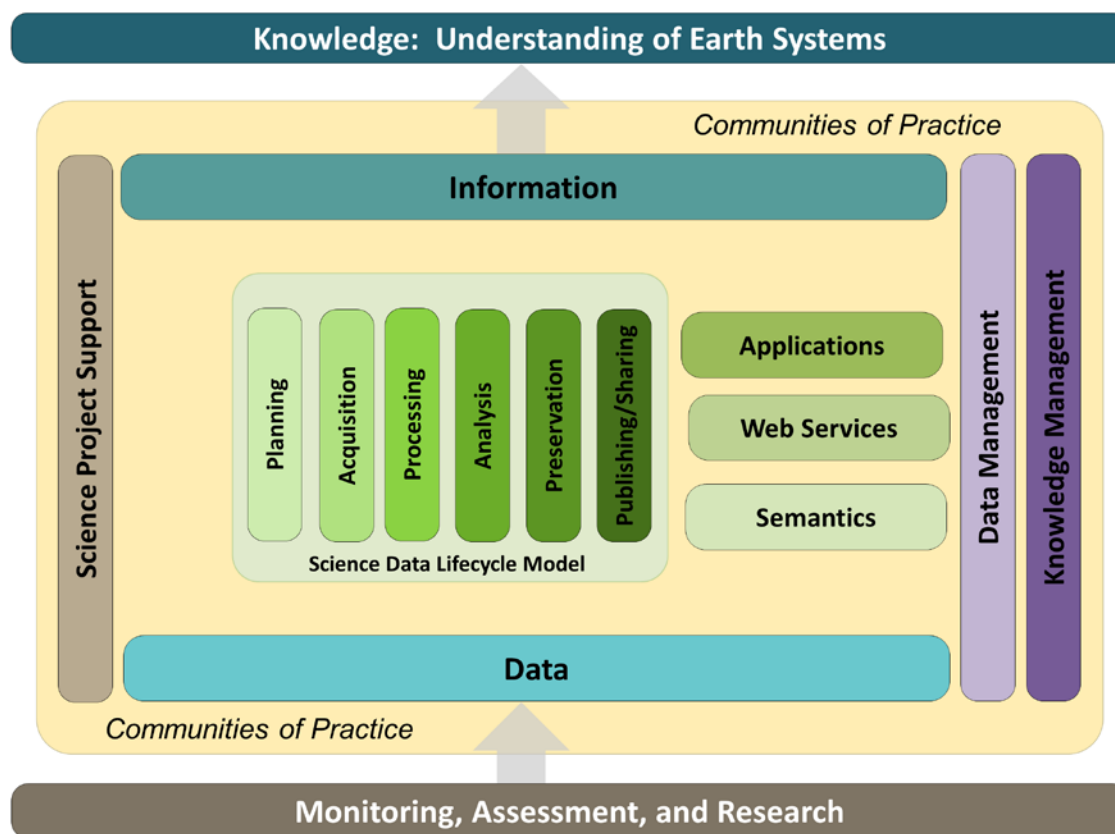


Figure 1. Community for Data Integration Science Support Framework (Faundeen and others, 2013).

The horizontal elements in the SSF (fig. 1) represent the “what” of the CDI: products and tools, and the mechanisms that mediate and contribute to the discovery and effective use of scientific data in systematic research. Data assets are managed within the context of the individual science projects, flowing horizontally from science project support through the Science Data Lifecycle (SDLC) processes, applications, and ultimately to data and knowledge management.

Monthly Forums and Annual Workshop/Webinar Series

The CDI monthly forums bring together the community to encourage discussion through presentations from USGS programs and science centers, as well as outside organizations, to share application development and new technologies, discuss data initiatives, and address data and metadata management challenges; furthermore, the CDI hosts an annual meeting or webinar that brings together scientists, managers, data managers, information technologists, and data users to leverage existing tools in support of mission-critical science and stimulate future thinking for emerging “big data” and other pertinent questions. Participants share activities and accomplishments through presentations, workshops, papers, poster sessions, and informal gatherings. These annual meetings and webinars have also become an outlet to share the results and progress of CDI-awarded projects. Together, the monthly forums and annual workshop/webinar series provide a key component in the networking and sharing among community members.

Community for Data Integration Monthly Forums

In fiscal year (FY) 2014, the CDI held monthly forums that covered a range of topics from USGS science and data activities to data lifecycle tools and technical solutions to data challenges from USGS science centers as well as external organizations (table 1).

Table 1. List of 2014 Community for Data monthly forums.

[FGDC-CSDGM, Federal Geographic Data Committee's Content Standard for Digital Geospatial Metadata; ESRI, Environmental Systems Research Institute; CDI, Community for Data Integration; OSTP, Office of Science and Technology Policy; OMB, Office of Management and Budget; ISO, International Organization for Standardization; USGS, U.S. Geological Survey; NOAA, National Oceanic and Atmospheric Administration; NWHC, National Wildlife Health Center]

Date	Presentation title	Speaker(s)
December 11, 2013	Metadata Wizard—An easy-to-use tool for creating FGDC-CSDGM metadata in ESRI ArcDesktop	Drew Ignizio, USGS Fort Collins Science Center.
February 12, 2014	Earth system informatics session of the 2013 CDI webinar series (For more information see the CDI 2013 Annual Report.)	Nancy Sternberg, Office of Administration and Enterprise Information; Tim Woods and Karen Armstrong, Office of Communications and Publishing; Carolyn Reid, Office of Science Quality and Integrity; and Mike Frame, USGS Core Science Analytics and Synthesis.
March 12, 2014	The DMPTool: Current status & future directions	Carly Strasser, California Digital Library.
April 9, 2014	National Land Cover Database Evaluation, Visualization and Analysis (NLCD EVA) Tool	Collin Homer and Chris Barnes, Earth Resources Observation and Science Center.
May 14, 2014	Data release on the Web: Meeting machine readability requirements for data (pursuant to OSTP and OMB)	Fran Lightsom, Viv Hutchison, Keith Kirk, John Faundeen, Carolyn Reid, Andrea Ostroff, Greg Miller, Greg Gunther, CDI Data Management Working Group and Peter Fox, Rensselaer Polytechnic Institute.
	Portable ISO 19115-2 Open Source Developer's Toolkit	Stan Smith, USGS Alaska Science Center and Josh Bradley, Fish and Wildlife Service Arctic Landscape Conservation Cooperative.
June 11, 2014	Mining the USGS data landscape	Lance Everette, Fort Collins Science Center.
July 8, 2014	CDI Workshop and Webinar at the ESIP Summer Meeting	See CDI Workshop and Webinar 2014.
August 13, 2014	Big data—Applications and use cases	Jeff Needham, Scale Abilities.
	NOAA data management activities	Jeff de La Baujardiere, National Oceanic and Atmospheric Association
	Implementation of the NWHC data and information management plan	Kim Miller, Jennifer Buckner, and Kathy Wesenberg, USGS National Wildlife Health Center.
September 10, 2014	Accelerating science through applied research computing	Jeff Falgout, USGS Core Science Analytics, Synthesis, and Libraries.
	Evaluation of downscaled general circulation models output	Lauren Hay, USGS National Research Program and Andy Bock, USGS Colorado Water Science Center.

2014 Community for Data Integration Workshop and Webinar

The CDI held a workshop at the Federation of Earth Science Information Partners' (ESIP) 2014 summer meeting in Frisco, Colorado, on July 8, 2014. The CDI members who did not attend the ESIP summer meeting in person were able to participate in the 2014 CDI Workshop via a WebEx webinar at no cost. About 40 participants were able to attend the workshop in person, and more than 100 participants were able to join remotely. The three and one-half hour workshop featured presentations from six CDI projects that received funding in 2014. The 2014 CDI Workshop provided an opportunity for CDI project awardees to share their projects with the community and get feedback for their ongoing work. Each presentation is described below.

Characterization of Earthquake Damage and Effects Using Social Media Data.—Michelle Guy, USGS National Earthquake Information Center, provided a technical overview of the USGS Twitter-based earthquake detection system and discussed the potential for rapid characterization of earthquake damage and effects using the 22 million “earthquake” tweets that the system had so far amassed. She also discussed the potential use of other social media sources such as Instagram for rapid impact assessment. Additionally, she described her project team’s plan to establish a data feed of the tweet-based earthquake detections to enable them to be shared and integrated with other derivative products.

Online, On Demand Access to Coastal Digital Elevation Models (DEMs).—Joseph Long, USGS St. Petersburg Coastal and Marine Science Center, gave an overview of his team’s project to integrate land and water-based elevation data sources by developing an online, geoprocessing service. He described how seamless data surfaces spanning the terrestrial-marine boundary will be available through a map-based Web interface, which implements the interpolation algorithms.

NASWeb Application Programming Interface (API)—Web Services Access to the Nonindigenous Aquatic Species (NAS) Database.—Matt Neilson, USGS Southeast Ecological Science Center, described his team’s work to create a publicly accessible Web Services API, which will provide an automated, machine-readable access point to the USGS NAS occurrence data. He explained how the open API will increase exposure for this integral USGS data resource, enhance its integration with other internal and external data sources, and stimulate novel applications and analysis of NAS occurrence data.

Summarization of National NEXRAD Data for Use in Biological Applications.—Tara Chesley-Preston, USGS Northern Rocky Mountain Science Center and Montana State University, reviewed her team’s project to create a proof-of-concept to advance methods of processing, summarizing, and distributing biological radar data of flying animals in a form more accessible and meaningful to the larger biological science community. She explained her team’s collaboration to develop algorithms to automate the collection and processing of NEXRAD datasets.

North American Bat Data Integration.—Laura Ellison, USGS Fort Collins Science Center, presented an overview of her team’s project to integrate two important bat-related datasets into an online Web application called the USGS Bat Population Data (BPD) Project so that they will be available to bat researchers and can inform conservation efforts for bats. The first dataset is the white-nose syndrome (WNS) diagnostic dataset, which can be used to better understand the distribution and spread of WNS, as well as the fungus that causes the disease. The second dataset is the Bat Banding Program files.

Adopt a Pixel—Data Infrastructure.—Ryan Longhenry, USGS Earth Resources Observation and Science Center, provided an overview of the CDI project to create a user upload capability, a data ingest process, and a data management system for the storage of imagery and metadata for the Adopt a Pixel system. He also explained how the project will enable incorporation of the Adopt a Pixel system into the USGS Earth Explorer Web-based user interface for query and download of the images by the end user.

Working Groups and Focus Groups

As part of networking and sharing of knowledge and experience, the CDI has expanded its community approach through working groups that form around common interests, help address challenges, and identify solutions that enable data integration efforts. Working groups have one or more leads and membership is open to all USGS employees, employees from other government agencies, and partners from external organizations. Some working groups have further subdivided into smaller focus groups that address more specific issues and report to their respective working group (table 2).

Table 2. Community for Data Integration 2014 working groups and focus groups.

[N/A, not applicable]

Working groups	Focus groups	Leads
Citizen Science	N/A	David L. Govoni
Connected Devices	N/A	Lance Everette and Tim Kern
Data Management	Science Data Lifecycle Model Team Data Policy Writing Team Data Management Web site Team Exit Survey Working Group Data Release Team Digital Object Identifiers (DOIs) Team	Heather Henkel and Viv Hutchison
Semantic Web	N/A	Janice M. Gordon and Frances L. Lightsom
Technology Stack	N/A	Daniella D. Birch and Roland J. Viger

Citizen Science Working Group

Citizen science is an innovative approach to advancing the Earth sciences that is being explored within the USGS and the Federal community at large. The purpose of the Citizen Science Working Group (CSWG) is to do the following:

- promote an understanding of the role and potential benefits of citizen science for USGS research;
- facilitate and enhance connections between the USGS and the larger Federal and public citizen science communities;
- provide access to information and tools to support the proper, effective, and creative use of citizen science data within the USGS;
- engage the public in USGS and partner science; and
- foster and improve the public's scientific literacy.

Because of the interest in citizen science at the Federal level, several working group members collaborated with other Federal agencies to help plan and support the creation of the Federal Community of Practice on Crowdsourcing and Citizen Science (FCPCCS) (U.S. EPA, 2015) and participate in its initial activities. In FY 2014, the CSWG focused all its effort on these broader and more collaborative activities as opposed to working as an isolated unit in the world of citizen science.

Citizen Science Working Group Accomplishments

As part of the USGS's engagement with the FCPCCS, members of the CSWG participated as planners, coordinators, or contributors in the activities and studies sponsored by the FCPCCS in collaboration with the Commons Lab of the Woodrow Wilson International Center for Scholars, including the following:

- A preliminary study entitled "An Exploratory Study on Barriers," which sought to understand the accelerators of and internal and external barriers to Federal use of crowdsourcing and citizen science (WCSTIP, 2014), and
- New Visions in Citizen Science, a November 2013 workshop that showcased federally sponsored citizen science and open innovation projects with the goal of examining different models of public participation. Special emphasis was placed on exploring project challenges, and the positive effects these projects can have on scientific literacy, research, and public policy. The resultant report highlights 17 case studies, including 3 from the USGS (Nature's Notebook, Did You Feel it?, and Tweet Earthquake Dispatch) (Bowser and Shanley, 2013).

Connected Devices Working Group

To help information technology (IT) professionals, software developers, and scientific researchers come to grips with the possibilities, opportunities, and challenges of the Internet of Things (Atzori and others, 2010), the CDI established the Connected Devices Working Group (CDWG). This group explores the use of cutting edge mobile tools, frameworks, and security considerations to support scientists. Some of the technologies scheduled for evaluation include wearable computing devices, cellular-based tracking devices, and mobile science support tools.

This working group supplants the legacy CDI Mobile Applications Working Group. Many of the ideas, functions, and efforts explored by the CDI Mobile Applications Working Group have been adopted by enterprise efforts, especially in the area of responsive design and mobile development frameworks. Although the CDWG continues to explore mobile application frameworks and technical advances, it also now delves into new topics associated with all-things Internet accessible.

Connected Devices Working Group Accomplishments

The CDWG works on multiple fronts: organizing group calls, coordinating with DigitalGov, lobbying Department of the Interior (DOI) and USGS policy contacts, and developing one-on-one contact with USGS staff considering projects in the realm of connected devices. This last effort focuses on developing capabilities that can be shared with the working group at large. Thanks to efforts by working group members, the USGS has done the following:

- Established a mobile application test framework,
- Implemented an identity management model that can be used by software developers, and
- Reconsidered the USGS Mobile Framework in light of new technical and policy directions.

Connected Devices Working Group Meetings and Presentations

The following are CDWG meetings and presentations:

- Cell-based Species Tracking Case Study–California Condor, Bob Waltermire [USGS Fort Collins Science Center (FORT)]
- Working with Streaming Data, Jeff Allen and Dell Long (USGS-FORT)
- Esri's Open-Source Tools for Geospatial Exploration, Andrew Turner (Esri)

- Refactored iPlover Application, Megan Hines [USGS Center for Data Analytics (CIDA)]
- The Digital Services Innovation Center Open and Structured Content Models, Lance Everette (USGS-FORT)

Data Management Working Group

The Data Management Working Group (DMWG) was established in 2010 to identify mechanisms for incorporating data management into USGS science activities and develop ways to educate scientists on the value of data management. Since its inception, the DMWG has continued to grow each year and now has a membership of about 100 people from across the USGS representing each mission area and region.

In FY 2014, the DMWG continued working on a variety of data management-related tasks while engaging the data management community with monthly meetings. Presentations at the monthly DMWG meetings included the following:

- Guidance on the Selection and Appraisal of Geospatial Content of Enduring Value, Colleen Cahill (Library of Congress)
- Managing Madness: Data Management in the Energy Resources Program, Greg Gunther (USGS)
- A Model Verification Process and Model Archive Repository, Martha Watt and Steve Tessler (USGS)
- Records Management: Initiatives, Opportunities, and Assistance, Chris Bartlett (USGS Records Manager)
- Duplicated Water Data—Causes, Implications, Solutions, Dorrie Gellenbeck (USGS)
- National Geospatial Data Asset Management Plan, Gita Urban-Mathieux (USGS)
- Data Management Approach for the Great Lakes Science Center, Scott Nelson (USGS)

Monthly updates to the CDI regarding DMWG activities were provided to the CDI Monthly Forums, and coleads participated on the CDI Coordinators group on a regular basis.

Data Management Working Group Accomplishments

The DMWG works by dividing into volunteer-led small groups to address topics of current interest and concern for the USGS. In 2014, the DMWG produced several products, best practices, and guidance based on the SDLC.

Data Management Web site

The data management Web site took on increased visibility as the new data management policies were developed for the USGS by supporting the content of the developing policies with guidance and best practices for compliance with the new directives (U.S. Geological Survey, 2015b). The Web site was continuously updated to reflect new best practices and guidance for compliance with the anticipated policies (fig. 2).

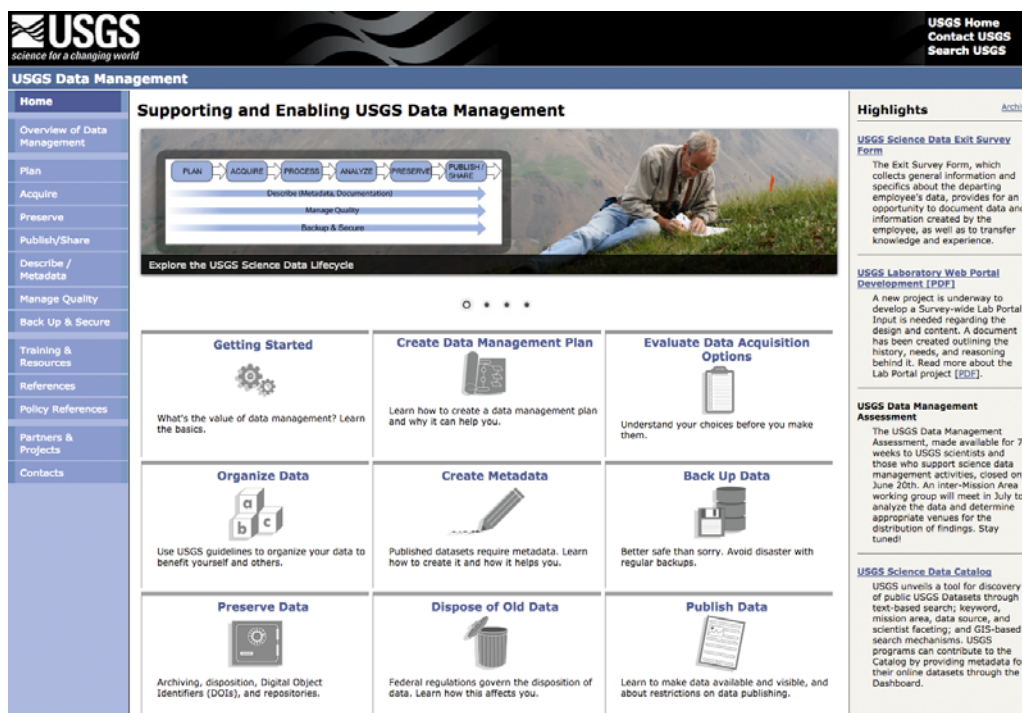


Figure 2. U.S. Geological Survey (USGS) data management Web site home page.

Science Data Lifecycle Model

The DMWG developed and published an SDLC model for the USGS upon which content for the data management Web site is based (Faundeen and others, 2013). The SDLC model illustrates the stages of data management within a project workflow, from planning for data management and acquiring data through preserving and sharing the data. The SDLC also includes cross-cutting, supporting activities like documenting the SDLC, managing quality, and backing up data.

USGS Science Data Exit Survey Form

The DMWG formed a team to assist programs and centers with tracking the status of valuable data and research products when scientists and other USGS staff depart. The team produced an Exit Survey to address this need and improved upon existing processes. Results of the effort are on the data management Web site (U.S. Geological Survey, 2015c).

Data Policy Writing Team

The Data Policy Writing Team was concentrated on writing four new data management policies for the USGS to be released as Instructional Memos (U.S. Geological Survey, 2015d):

- Scientific Data Management Foundation, 2015-01
- Metadata for Scientific Data, Software, and Other Information Products, 2015-02
- Review and Approval of Scientific Data for Release, 2015-03
- Preservation Requirements for Digital Scientific Data, 2015-04

The Policy Team also developed and published guidance on identifying interpretive versus noninterpretive data to support understanding and proper interpretation of the policies (U.S. Geological Survey, 2014a). A decision-tree was also created (U.S. Geological Survey, 2014b).

Data Release Team

Another team in the DMWG, the Data Release Team, continued efforts on developing guidance for data release in the USGS. The Data Release Team produced a graphic and text-based workflow to guide scientists on how to release data to the Web. Additionally, the group produced guidance documents focusing on techniques for good review of metadata and data (U.S. Geological Survey, 2015e).

Digital Object Identifiers (DOIs) Team

Finally, the DMWG established a small team to focus on DOIs for the USGS. The outcome was twofold: a list of requirements for enhancements to the DOI Tool, which was established as a resource for scientists to create and apply DOIs to data, and a Frequently Asked Questions document (the document is accessible only to internal USGS employees) (U.S. Geological Survey, 2014c).

Data Management Working Group Meetings and Presentations

Members from the DMWG gave the following presentations regarding data management to external groups in FY 2014:

- Data Stewardship—Best Practices and the USGS Experience, Hutchison, V., Bureau of Reclamation, Denver Colorado, 2014;
- Engage in Culture Change—Fundamentals for Working in a Data Centric World, Hutchison, V., Kathleen A. Zar Symposium, University of Chicago, 2014;
- Data Management Successes, USGS Ecosystems Seminar, Henkel, H., 2014;
- Data Management in the USGS—Everything You Wanted to Know But Were Afraid to Ask, USGS National Research Program, Henkel, H., 2014; and
- USGS Data Management Web site and Training, UNESCO-IHE International Students Partnership Lectures, Henkel, H., 2014.

Data Management Working Group Feedback from Stakeholders

“As I indicated at the Open Data Meeting last week, I believe that your data management modules are the single best contributions to DOI data stewardship that have been produced thus far. In my view, DOI needs to work from the ground up to change agency culture and make data management a permanent part of the scientific method. Your modules do just that. Thank you for developing them.”

— Doug Clark, Bureau of Reclamation

“This is really a nice Web site!! <http://www.usgs.gov/datamanagement/index.php> Congrats!”

— Bob Cook, Oak Ridge National Laboratory

“Thanks for all your hard work on this great informational resource.”

— Keith Kirk, USGS Office of Science Quality and Integrity

Semantic Web Working Group

The main goal of the Semantic Web Working Group (SWWG) is to explore Semantic Web technologies as a possible solution for integrating and exposing related but disparate datasets that are commonly only accessible through one application "silo" at a time. The SWWG explores these solutions by examining different real world use cases where scientists have difficulty finding usable data to answer research questions. One of the main functions of the SWWG is to share knowledge and build on each other's experiences.

Semantic Web Working Group Accomplishments

In 2014, the SWWG sponsored a project to develop a prototype vocabulary server that can be used to incorporate the USGS Thesaurus and Biocomplexity Thesaurus into metadata creation tools and the interfaces of metadata catalogs such as the USGS Science Data Catalog. Initial products from this project are a set of use cases that describe how the vocabulary server would respond to the other systems (fig. 3) (Lightsom, 2014). The vocabulary server provides the services that allow the metadata tool, catalog manager, and catalog user interface to offer controlled vocabularies for inclusion in metadata records and for assistance in discovering those records in a data catalog.

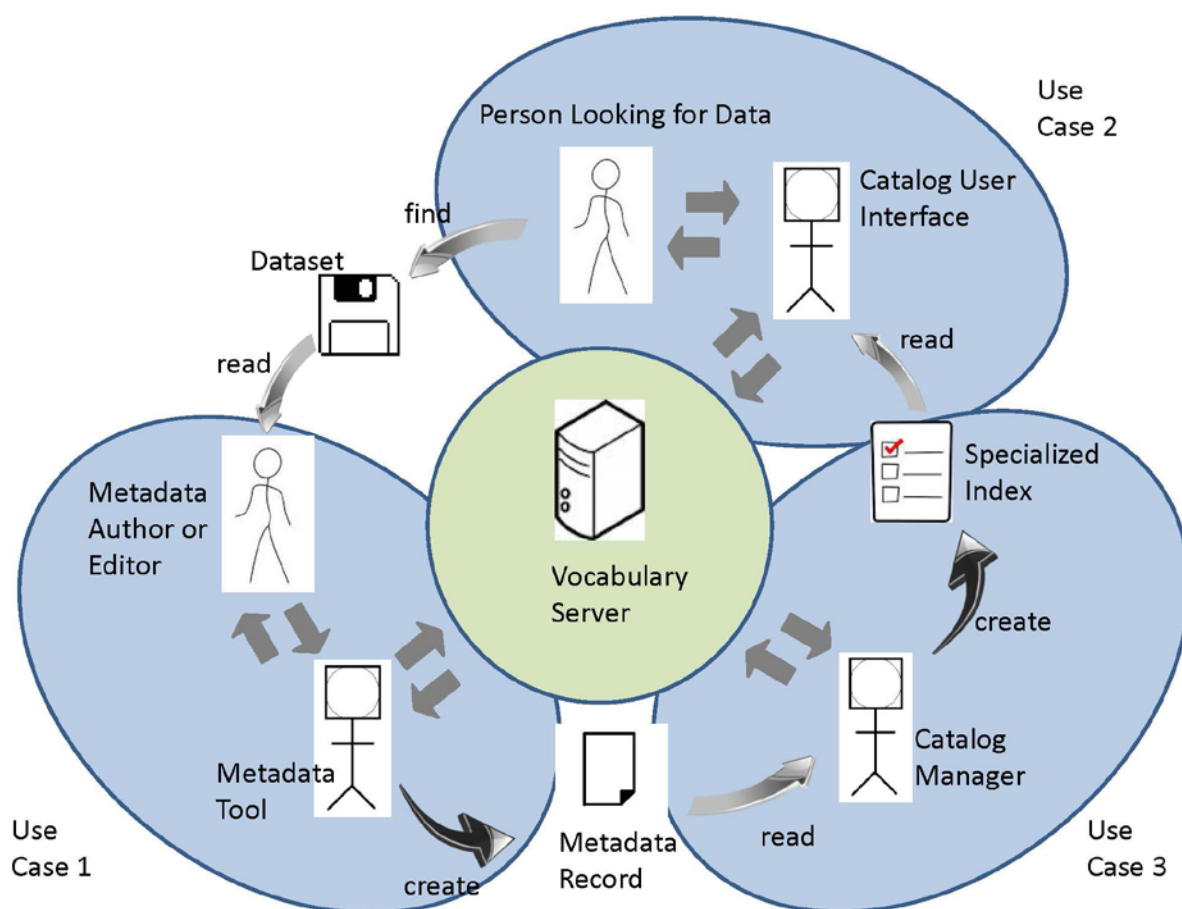


Figure 3. A diagram showing the three controlled vocabulary use cases and their relationships to each other.

Semantic Web Working Group Meetings and Presentations

The SWWG hosted the following meetings and presentations in FY 2014:

- Vocabulary Management at the Biological and Chemical Oceanography Data Management Office (BCO-DMO), Dicky Allison, Woods Hole Oceanographic Institution, December 12, 2013.
 - The BCO-DMO is funded to serve the data management requirements of investigators funded by the U.S. National Science Foundation (NSF) Biological and Chemical Oceanography Sections and the Division of Polar Programs Antarctic Sciences Organisms and Ecosystems Program. Variable names supplied by contributing scientists are matched to BCO-DMO standard names, which are in turn mapped, if possible, to the terms used by the SeaDataNet community served by the Natural Environment Research Council Vocabulary Server at the British Oceanographic Data Centre. BCO-DMO is currently (2015) exploring ways to also incorporate or map other controlled vocabularies, including the Darwin Core terms, into the dataset metadata and our evolving ontology.
- What do we mean by Semantic Web?, Informal group discussion, January 9, 2014.
- EarthCube Roadmap for Creating the Semantic/Ontologic Infrastructure for the Geosciences, Informal group discussion, February 13, 2014.
- Why We Need a Semantic Web Framework for Marine Ecosystem Indicators, Stace Beaulieu, Woods Hole Oceanographic Institution, March 13, 2014.
 - Ecosystem-based management of Large Marine Ecosystems involves the sharing of data and information products among a diverse set of stakeholders—from environmental and fisheries scientists to policy makers and the public. Tracing a synthesized data product back to its original data sources is important for verification, understanding, and comparing with similar data products or indicators monitored by different stakeholder groups; and can be done by using the W3C provenance Linked Data standard and linking to domain-specific ontologies to give meaning to the source datasets and derived products.
- Ontologies for Emergency and Disaster Management (The application of geospatial linked data), Stephane Fellay and John Davidson, Image Matters, LLC, April 10, 2014.
 - This work is using linked data concepts/standards to integrate two building-block ontologies that have been defined in cooperation with Open Geospatial Consortium (OGC) ("Core Spatial Ontology" and "Core Incident Ontology for E&DM") to demonstrate services and capabilities for semantic integration/interoperability. The real take away is about mechanisms, patterns, and best practices for using partitioned ontologies as building blocks for producing linked data.
- How are semantic techniques becoming part of your toolkit?, Informal group discussion, May 8, 2014.
- Behind the curtain at the Global Change Information System, Peter Fox, Rensselaer Polytechnic Institute's Tetherless World Constellation (RPI TWC), June 12, 2014.
 - When the U.S. global change report was released in May 2014, it was accompanied by an extensive information system providing supporting data. This Global Change Information System works with Semantic Web technology.
- Informal progress report from the SWWG Vocabulary Services Use Case Team, August 14, 2014.

- Informal group discussion of the Wikipedia article on Semantic Web, especially Section 4—Standards (which has a clean little diagram and many links), September 11, 2014. Questions discussed at the meeting include the following:
 - Are there standards on this list that would be useful for USGS data integration efforts if we only knew how to use them?
 - Might we think about sponsoring training for USGS on some of these standards?

Technology Stack Working Group

The purpose of the Technology Stack Working Group (TSWG) is to build a community of scientists, data managers, and technology practitioners who discover, evaluate, and share knowledge about the best available tools and practices related to scientific computing, data creation, and data publication with the goal of improving scientific workflows through data integration. There are a couple additional goals of this group. The first is to provide a unifying umbrella under which many member-defined technology focus groups can be hosted. The overarching working group seeks to provide coordination across the focus groups with the goal of developing a more holistic understanding of several technologies, standards, and protocols that are constantly evolving. The second is to help inform thinking in other working groups, such as the Data Management Working Group, by providing expertise on how concepts, protocols, Fundamental Science Practices (FSP), and the technical memoranda, could be implemented.

Technology Stack Working Group Accomplishments

A big accomplishment for TSWG was engaging a new coordinator, Daniella Birch. She has brought new topics to the TSWG and engaged members, including previous coordinators. The TSWG has plans to become more active in FY 2015 including hosting a new series of talks on topics such as Agency status of cloud computing and Web mapping.

Annual Community for Data Integration Request for Proposals

For the past few years, the CDI has provided funding support for projects that promote data integration and do the following:

- focus on short-term benefits to science,
- leverage existing capabilities,
- apply solutions/methodologies that can be replicated,
- ensure sustainability,
- seek substantial return on investment,
- expose corporate data,
- organize science models and outputs, and
- preserve and access project data.

The 2014 RFP implemented a new two-phased approach for soliciting and reviewing proposals in order to reduce the time and effort of submissions for applicants and involve the community in the review process (U.S. Geological Survey, 2013). The first phase is the statements of interest (SOIs) phase, which requests applicants to submit short, 2-page proposals summarizing their idea. The second phase is the Full Proposals in which a set of SOIs selected with the input of the community are invited to submit full proposals that fully outline the project goals, deliverables, and proposed budget. The

improvement of this year's process enabled the CDI to address the concerns and feedback obtained from the 2013 RFP experience.

Phase I—Statements of Interest

A formal RFP was announced on September 11, 2013, calling for SOIs in four different CDI Categories that support the SSF:

Category 1—Management, Policy, and Standards.—Data management includes data and metadata standards and policies, and is necessary in all phases of the science data life cycle (SDLC) from planning and data acquisition through scientific research to finished information products. Knowledge management involves the creation, standardized documentation, and organization of artifacts describing or encapsulating knowledge resulting in the creation of reusable knowledge bases. Activities may involve the development, application, and (or) testing of data integration processes, protocols, and products that are or result in improved or more effective data integration management, policies, and standards.

Category 2—Computational Tools and Services.—Computational tools and services include SDLC processes, tools, and services that move data through the SDLC, relate human and machine interactions, and interact with data through technology. Activities are primarily technical in scope and aimed toward the development of applications, Web services, semantics, or combinations of technologies centered upon the advancement of data discovery and integration.

Category 3—Data and Information Assets.—SDLC processes, data management, and knowledge management processes operate on data and information (assets). Activities may be aimed at creating or improving data and information products. They may involve or result in the development, application and (or) testing of data and information assets that support, facilitate, or improve data integration and the transition from data to information to knowledge.

Category 4—Community Innovation.—Community innovation may involve or result in the development, application and (or) testing of unique data and information assets, tools, or services; and demonstrate novel, innovative approaches and solutions for data integration; or the development of new unique tools and services for data management and integration.

The SOI phase was met with great enthusiasm and a mixed spread of SOIs were submitted to each CDI SSF category (table 3). A total of 64 SOIs were submitted for FY 2014 (table 3).

Table 3. Number of proposals submitted for the Community for Data Integration fiscal year 2014 request for proposals by Science Support Framework category.

[SSF, Science Support Framework]

Category number	SSF category name	Number of proposals submitted
Category 1	Management, Policy, and Standards	7
Category 2	Computational Tools and Services	16
Category 3	Data and Information Assets	30
Category 4	Community Innovation	11
Total:		64

In order to increase transparency of the process and participation from the community, members were invited to evaluate SOIs in each category and vote for proposals that met the CDI goals and SSF through an online voting event. Members were able to leave comments on the voting Web site for the

SOI applicants to address. For each category, the CDI held open SOI review meetings to discuss the SOIs and determine voting results. These review meetings provided a forum for SOI applicants to answer questions posed by the community about their proposal. At the end of SOI review session, the voting was closed, and the community viewed the voting results. Participants as a group agreed on several SOIs in each category that would be recommended to move forward into the Full Proposal Review stage.

The CDI Coordinators also reviewed the SOIs and provided feedback for final recommendations. With the Executive Sponsor's approval, 21 SOIs were selected to be invited to submit full proposals. Two proposal authors chose not to submit full proposals; therefore, a total of 19 full proposals were submitted to be reviewed in the second phase.

Phase II—Full Proposals

The CDI convened a CDI Review Panel charged with reviewing the 19 full proposals and providing recommendations for funding to the Executive Sponsor. All panel reviewers were USGS Federal employees who volunteered their time. The reviewers encompassed a wide variety of expertise and represented a broad range of mission areas, regions, and programs. Most of the reviewers were not affiliated with the CDI and participated out of interest in the topic of data integration. In order to maintain objectivity and anonymity, the reviewers disclosed potential conflicts of interest by signing a conflict of interest form and were asked not to divulge the identity of their CDI Review Panel peers during the process.

Review Panel Process

Reviewers were assigned to lead the discussion of 2–3 proposals and were responsible for an indepth knowledge of these proposals. All reviewers individually reviewed and documented a summary of the strengths and weaknesses of each proposal based on guidance from the RFP Document. Reviewers also scored each proposal based on the required weighted sections as follows:

- Scope (25 percent)
- Technical approach (25 percent)
- Project experience (25 percent)
- Commitment to effort (15 percent)
- Budget (5 percent)
- Timeline (5 percent)

Having read and evaluated the proposals individually, the reviewers then participated in several group review panel sessions to discuss the merits of each proposal. Reviewers discussed the strengths and weaknesses of each proposal and gave their recommended scores. Scores for each proposal were averaged to obtain an initial score for the proposal. During a final review session, the panel collectively discussed the proposal rankings and agreed upon a final recommendation for each proposal. Six projects were recommended for funding by the CDI Review Panel (table 4).

Bureau-Wide Applications

A notable outcome of the CDI Review Panel was the identification of three proposals as having Bureau-wide applications (table 4). Although the three proposals were written with a specific scope of work, the CDI Review Panel felt that the proposals could be implemented with a broader perspective to serve the needs of many other USGS mission areas, regions, centers, and programs; therefore, the CDI

Review Panel asked that the proposals be given additional funding for a second year in order to make their activities more applicable across the USGS. These Bureau-wide Applications projects were redeveloped to include broader collaboration and effect and will be funded for a total of 2 years (FY 2014–FY 2015).

Recommendations

The recommendations of the CDI Review Panel were submitted to CDI Executive Sponsor, Kevin Gallagher (Associate Director, Core Science Systems mission area) for final approval. On March 12, 2014, Kevin announced funding for the six projects and three Bureau-wide applications projects. The “CDI Projects” section describes the projects and their accomplishments in more detail. The 2014 RFP projects and products are also available online (U.S. Geological Survey, 2015f).

Table 4. Overview of the Community for Data Integration request for proposals projects funded in fiscal year 2014 (in alphabetical order).

Title	Principal investigator(s)	Program
Adopt a Pixel—Data Infrastructure	Ryan Longhenry and Eric C. Wood	Earth Resources Observation Systems
Characterization of Earthquake Damage and Effects Using Social Media Data	Michelle Guy and Paul S. Earle	Geologic Hazards Science Center
Geographic Searches for USGS Publications ¹	Katherine E. Wesenberg	National Wildlife Health Center
NASWeb API Web Services to the NAS Database	Pam Fuller and Craig Conselmann	Southeast Ecological Science Center & National Wetlands Research Center
North American Bat Data Integration	Laura E. Ellison	Fort Collins Science Center
Online Merging and Gridding of Topographic and Bathymetric Data Sources	Joseph W. Long and Sky Bristol	St. Petersburg Coastal and Marine Geology Science Center & Core Science Analytics, Synthesis and Libraries
Portable ISO 19115-2 Open Source Developer’s Toolkit ¹	Stan Smith and Joshua Bradley	Alaska Science Center & US Fish and Wildlife Service
Summarization of National NEXRAD Data for use in Biological Applications	Robert H. Diehl	Northern Rocky Mountain Science Center
Use of Controlled Vocabularies in USGS Information Applications: Requirements Analysis for Automated Processes and Services ¹	Frances L. Lightsom	Woods Hole Coastal and Marine Science Center

¹Bureau-wide Application Project

Community for Data Integration Projects

The FY 2014 projects represented many elements of the SSF including data, information, communities of practice, applications, Web services, semantics, acquisition, processing, and analysis. More than half the projects developed Web services as one of their deliverables and five projects developed new or improved existing applications to better document, integrate, expose, and improve access to data and information. The following sections provide details about the 2014 CDI projects listed in table 4 and their accomplishments.

Adopt-a-Pixel—Data Infrastructure

Adopt-a-Pixel – Data Infrastructure (AaP-DI) provides the basis for a new data acquisition system for ground-reference data. These data will be used to complement existing and future remote sensing collections by providing geospatially tagged ground-based landscape imagery and land cover of an exact location from six viewing aspects. The goal is for AaP-DI to enable citizen participation in Landsat science.

Deliverables

The deliverables for the Adopt-a-Pixel project were as follows:

- The Adopt-a-Pixel Web interface (fig. 4) is a data upload portal that allows citizen scientists to upload geospatially tagged ground-based landscape photographs.
- The uploaded photographs are run through a data ingest and archive script, which places the images in a managed state by creating a database record; and browse images, metadata, and an archived copy of the data.
- Scientists can query and download the Adopt-a-Pixel ground-reference data to verify land cover type through EarthExplorer, which is a pre-existing Web-based user interface that provides open access to the data.

The screenshot shows the USGS Adopt-a-Pixel web interface. At the top, there's a USGS logo and navigation links. The main content area is divided into several sections. On the left, there's a form for entering metadata: 'Land Cover Type' (a dropdown menu set to 'Rangeland'), 'Device Latitude' (a text input field with '36.45089'), 'Device Longitude' (a text input field with '-105.57519'), 'Date Taken' (a date input field with '2014-09-22'), 'Time Taken' (a time input field with '21:55:00'), 'Time Zone' (a dropdown menu set to 'UTC'), and 'Keywords' (a text input field with 'Forest Disturbance'). To the right of the form, there's a grid of six camera views showing different directions: North, South, East, West, Up, and Down. Each view shows a landscape with mountains and a dirt road. At the bottom right, there's a 'Submit' button.

Figure 4. The Adopt-a-Pixel Web interface shows uploaded imagery from a single point facing north, south, east, west, up, and down to give a clear view of landscape features in all directions.

Characterization of Earthquake Damage and Effects Using Social-Media Data

People in the locality of earthquakes are publishing anecdotal information about the shaking within seconds of their occurrences through social network technologies such as Twitter. In contrast, depending on the size and location of the earthquake, scientific alerts can take between two to twenty minutes to publish. The goals of this project are to assess earthquake significance, damage, and effects information as effects unfold by leveraging expeditious, free, and ubiquitous social-media data to enhance our response to earthquake significance, damage, and effects.

The project leverages an existing system that acquires data from Twitter of earthquake related tweets, and geocoding from Yahoo. The acquired data stream is archived into a Postgres database that is filtered and monitored by the detection application (Tedect). The social media data acquisition and distribution application (TED) was enhanced to concurrently feed the focused data stream into an Elasticsearch index. Elasticsearch is an open source tool built on Apache Lucene, a high-performance text search engine library, that creates optimized searchable data in JavaScript Object Notation (JSON) format that is in turn visualized by Kibana, a Web-based interactive interface. Elasticsearch and Kibana are new technologies that this project team was not aware of at the time the proposal was written, but proved to be a valuable, low cost way of providing rapid indications of earthquake significance, damage, and effects.

As a proof of concept, the geographic data associated with the project were upgraded to make them more readily and programmatically available for collaborators. The creation of geographic information systems (GIS) services for sharing and distributing data was deemed the easiest data feed

source to most readily integrate with a variety of GIS end users including the Federal Emergency Management Agency (FEMA). Implementation of two service types on a per event basis was established for manual configuration following significant, detected events. The first is a point-based tweet dataset that can be mapped and queried. The second is a raster-based heat map of tweet density, which covers the region of interest. Once manually configured, these services are created to acquire related data at three distinct time intervals after an event, which supports the series of post detection summary reports.

All this data acquisition, processing, analysis, and organization leads to the automated production of valuable earthquake significance, damage, and effects information for use by practitioners (scientists, earthquake responders, and collaborators). The rapid availability of such earthquake characterization information allows practitioners to customize their alerts from this system according to event significance.

The proposed project included integration of additional social media data sources; however, further analysis of data from Flickr and Instagram determined the signal to noise ratio too high to be of benefit in the scope of this project. Future efforts and tool development to more efficiently filter and assess the data may prove fruitful; however, with the resources available at this time, focusing on Twitter data and media proved more effective.

The outcomes from this project support USGS goals of advancing scientists' understanding of earthquake effects, establishing data sharing and collaboration, and providing additional rapid situational awareness.

Deliverables

The deliverables for the Characterization of Earthquake Damage and Effects Using Social-Media Data project were as follows:

- Updated the Social Media Earthquake Acquisition and Distribution Application, also known as TED. The enhancements to the system include derivation and distribution of earthquake significance summary reports and archival of tweet-based event detections. The code is available on a private USGS stash git repository. Efforts are still underway to migrate to a publicly accessible git repository, which is planned for completion in 2015.
- The Social Media Earthquake Detection Application, also known as Tedect, was updated to send tweet-based event detections into TED and the performance was improved. The code is on the USGS stash git repository. An older reviewed version is publicly available at <https://github.com/mguy-usgs/tedect>.
- Earthquake Significance Summary Reports are automatically derived for tweet-based event detections at 5 and 10 minutes after events take place (fig. 5). The reports give rapid indications of public interest in the event for use internally and by collaborators.
- Web Services
 - The project team established a workflow for using the Kibana Visualization and Data Analysis Tools to, in near real time, determine earthquake significance, damage, and effects as illustrated in figures 6 and 7. As of 2014, this open-source service has no free mechanism to secure the data in order to preserve data integrity, so it remains internal.
 - As a proof of concept, with coordination from FEMA, a GIS service feed for detected events was implemented. An example heatmap created from this service is included

in figure 8 and is available at

<http://geohazards.usgs.gov/arcgis/rest/services/NapaHeatmap/MapServer>.

- The Integrated Social Media and Seismic Earthquake Dataset provides validation of the system, and analysis demonstrates the system performance as described in table 5. The analysis of this dataset was shared publically at a scientific conferences including the American Geophysical Union (AGU) in December 2014 and the Association of Computing Machinery (ACM) Special Interest Group Knowledge Discovery and Data Mining (SIGKDD) Learning about Emergencies from Social Information (LESI) workshop in August 2014.

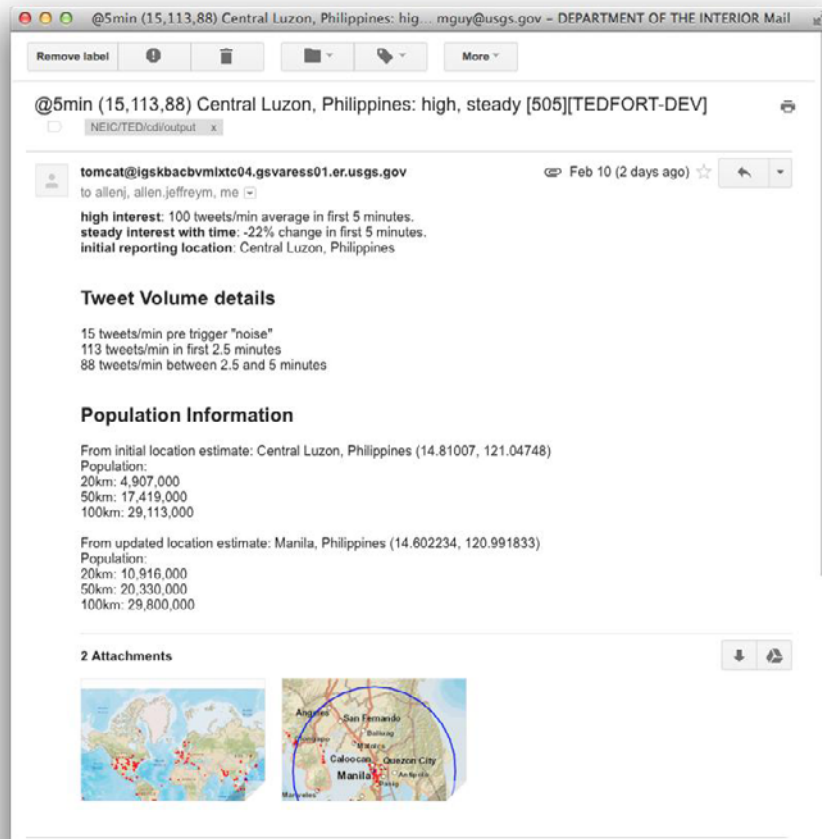


Figure 5. This figure shows the earthquake significance summary report that is automatically derived and emailed 5 and 10 minutes after the event detection.

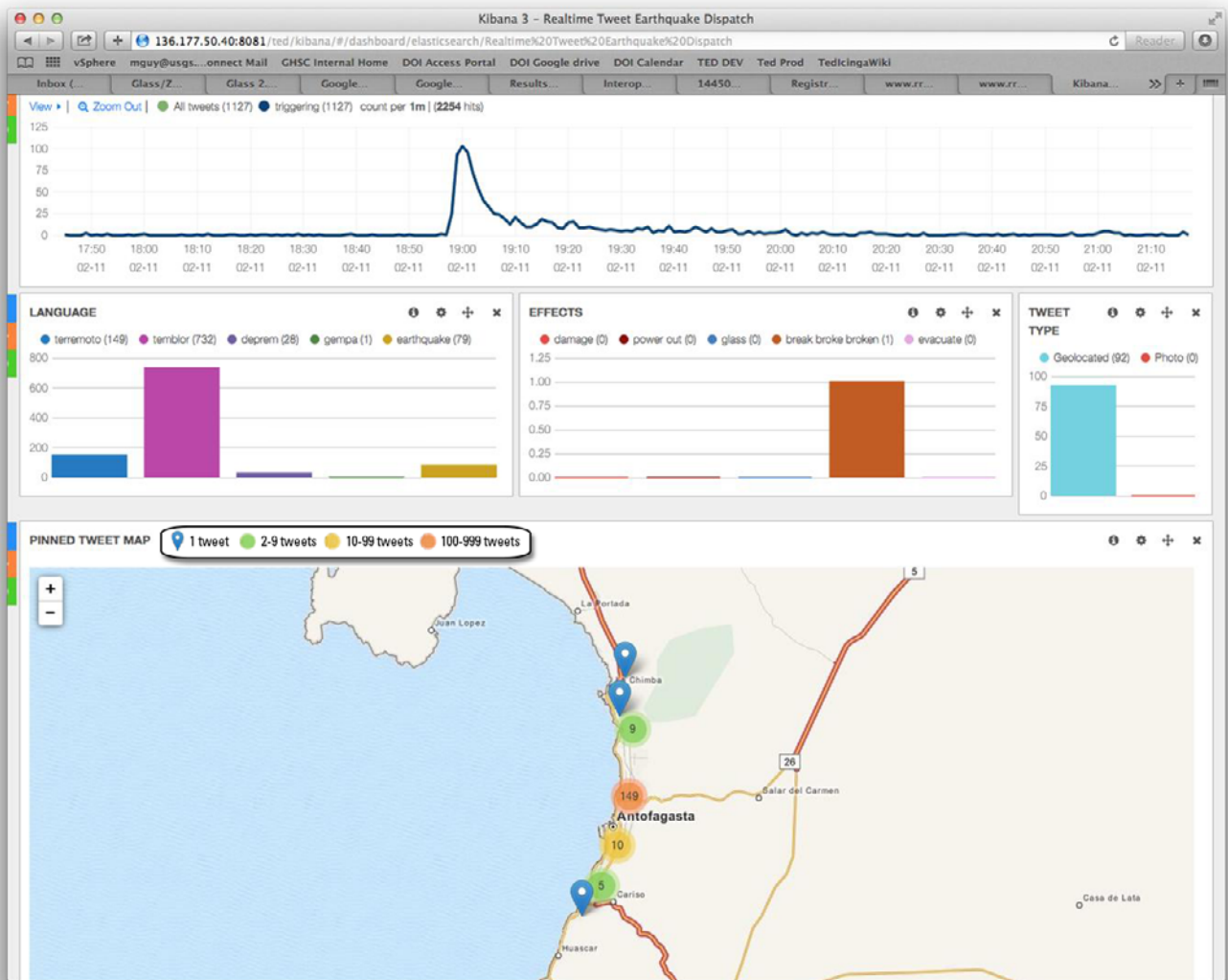


Figure 6. This figure shows a screenshot of the Kibana interactive visualization and data analysis tool which includes charts showing the event detection time based on a spike in the number of tweets over time (top), the language of the tweet (left center), keywords used that indicate the earthquake's shaking intensity and effects (center), and the type of tweet (right center). The tool's interface also includes a pinned tweet map, which shows the tweet locations and the density of tweets at those locations (bottom).

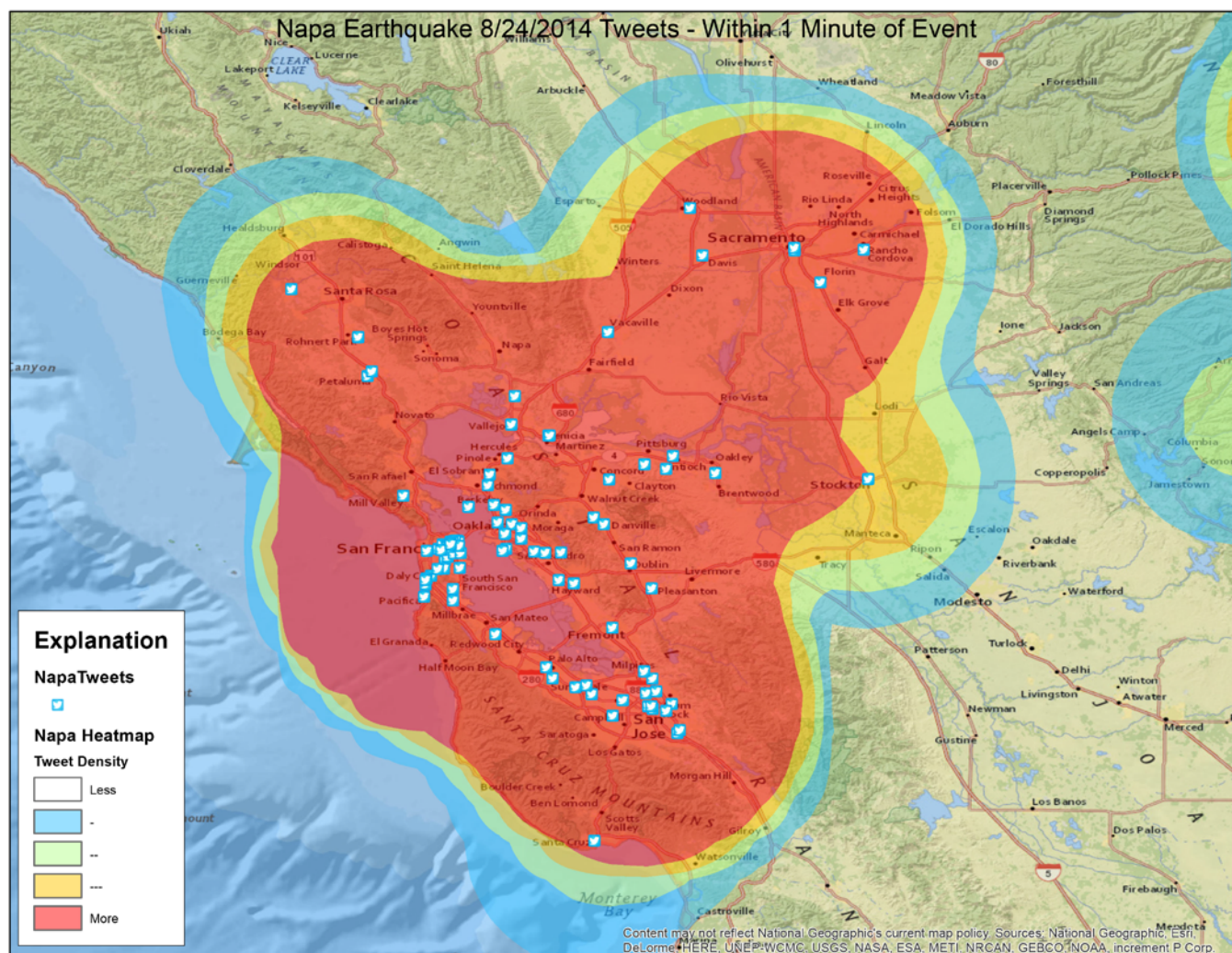


Figure 8. Heatmap produced from ArcGIS service shows the distribution and density reporting tweets.

Table 5. Analyses of the integrated social media based event detections and seismically derived events.
[% , percent; < , less than; km, kilometer]

Tweet-based event detection delay	50% < 1 minute	90% < 2 minutes
Tweet-based event location accuracy	65% within 100 km	90% within 200 km
Earthquake magnitude range for tweet-based event detection	Minimum 1.4	Maximum 8.2

Geographic Searches for USGS Publications

The purpose of this project is to improve the USGS Publications Warehouse so that a person can search for USGS publications by geographic region in addition to existing search criteria; for example, one could search using map zooms or congressional districts. The addition of geographic searches allows users to narrow their search results to specific areas of interest, which reduces the time required

to sift through all results outside the area of interest. The improved Publications Warehouse is planning to provide Web services so geographic searches for publications can be embedded in other Web sites. The project team plans to improve the Publications Warehouse software and database; develop software for reviewing publications and adding geographic polygons; and reuse or develop geographic polygons for initial regions of interest. The Coastal and Marine Geology Program (CMGP) is adding geographic polygons to roughly 10,000 publications that have CMGP authors and adding geographic polygons to new CMGP publications. The project team is developing workflows for adding geographic polygons to all new USGS publications. The team is exploring the possibility of adding geographic polygons to existing publications in a centrally prioritized, funded, and managed manner; or asking mission areas, regions, programs or science centers to take the lead in adding polygons for their publications.

Deliverables

The geographic searches for USGS publications project faced several challenges in FY 2014 that included departures of three original principal investigators (PIs) from the project. These dynamics hindered progress and challenged the effectiveness of the project's progress; however, the project progressed by accomplishing the following:

- Decided on technical implementation for ScienceBase and Pubs Warehouse to exchange information,
- Determined that ScienceBase Footprint Studio is sufficient without addition of primary technical capabilities (fig. 9), and
- Updated interface to allow users to search within a geographic area for publications and backend capabilities in PubsWarehouse to accept geospatial footprints related to publications (fig. 10).

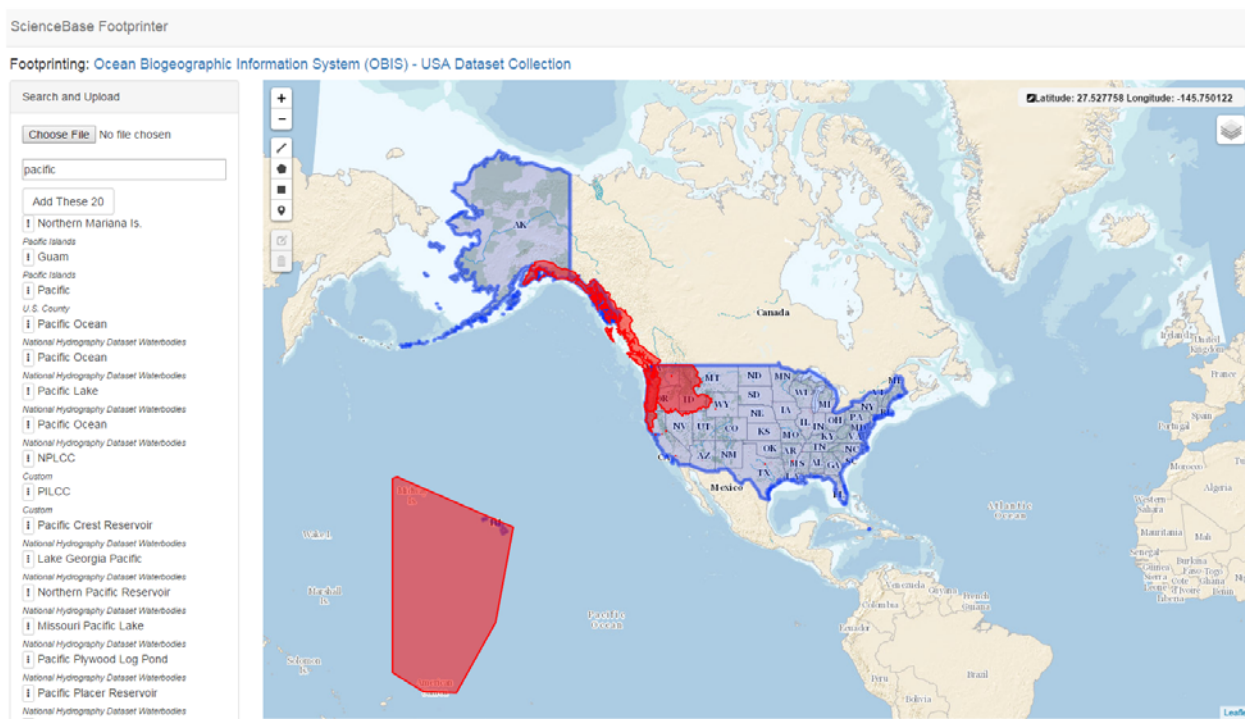


Figure 9. Screenshot of ScienceBase Footprint Studio displaying the footprint for the Ocean Biogeographic Information System (OBIS) – United States of America Dataset Collection.

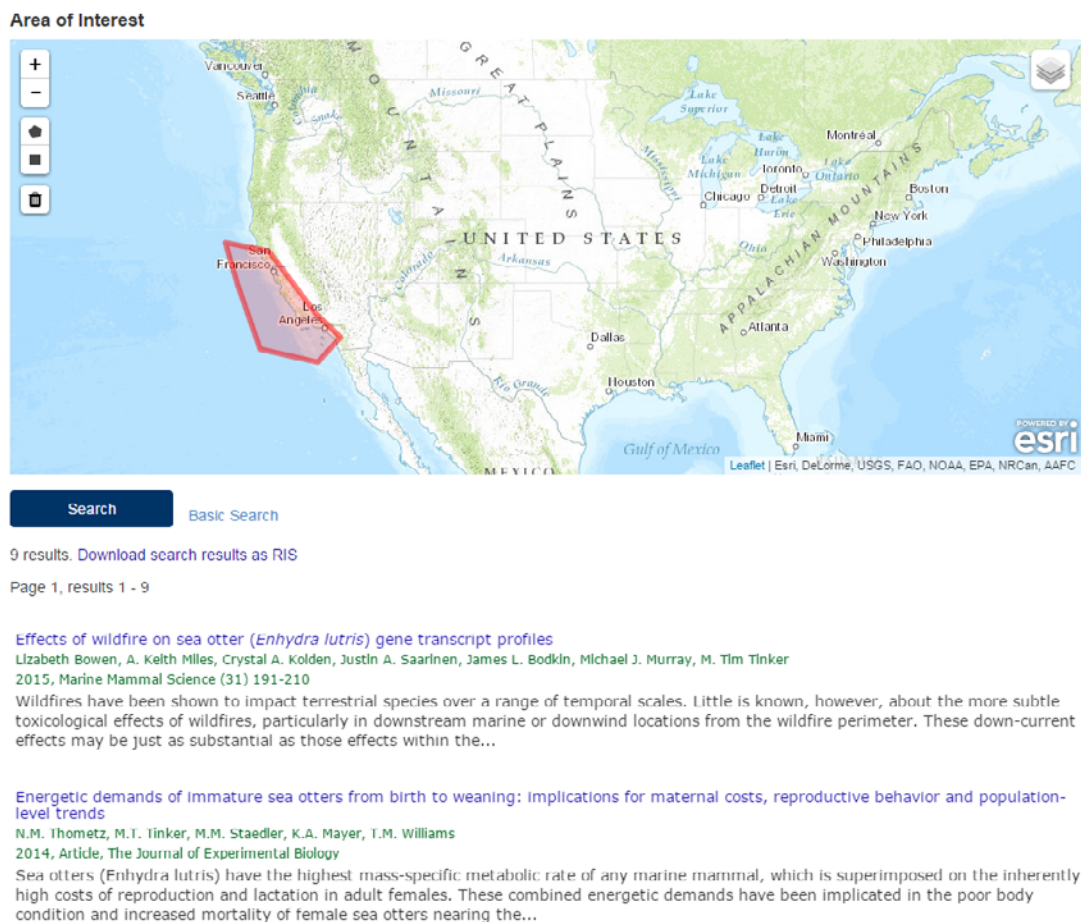


Figure 10. Screenshot of the PubsWarehouse advanced search function, which now allows geospatial footprint queries.

NASWeb Application Programming Interface—Web Services Access to the Nonindigenous Aquatic Species Database

The national Nonindigenous Aquatic Species (NAS) Database Program serves as a repository for georeferenced occurrence data on introduced aquatic organisms across the Nation. The NAS Program, including the database and Web site, is a well-known resource and has been widely referenced in peer-reviewed literature, agency reports, State and national management plans, news articles, and other locations (U.S. Geological Survey, 2015g). The NAS Database is used for analyzing risk, monitoring and prevention, managing and planning, and policymaking. As of 2014, access to NAS occurrence data is restricted to three channels: directly through the NAS Web site (primarily single occurrence records and species distribution maps), distributed databases [for example, Global Biodiversity Information Facility (GBIF) and Biodiversity Information Serving Our Nation (BISON)], and custom data queries directly through NAS Program staff. The goal of this project is to design, test, and implement a Web Services API for access to NAS occurrence data. This novel access method to the NAS database will increase exposure and accessibility of NAS occurrence data to a broader array of customers and allow for innovative uses of NAS data and integration with external data sources.

Deliverables

The deliverables for the NASWeb Application Programming Interface project were as follows:

- Web Services API for NAS data
The Web services API allows for access to information on occurrence data stored within the NAS database (fig. 11). The Web services API was designed and completed in September 2014. The API users can request information on a single specific occurrence record, or multiple records through query parameters. The API was designed to provide 2 levels of access: anonymous public access can retrieve 100 records per API request, and users possessing an API key (generated after discussion with NAS staff about scope of data use and potential data limitations) have unlimited access. The API was opened to the public on October 15, 2014. Public launch was delayed slightly from initial timeline because of slight complications in technical design and issues concerning handling and storage of personally identifying information (PII) associated with users granted an API key.
- NAS Web Services API Documentation
Documentation of the API has been generated, and includes information on the URL used for API requests, all potential query parameters, and the data format of query results (U.S. Geological Survey, 2015h).

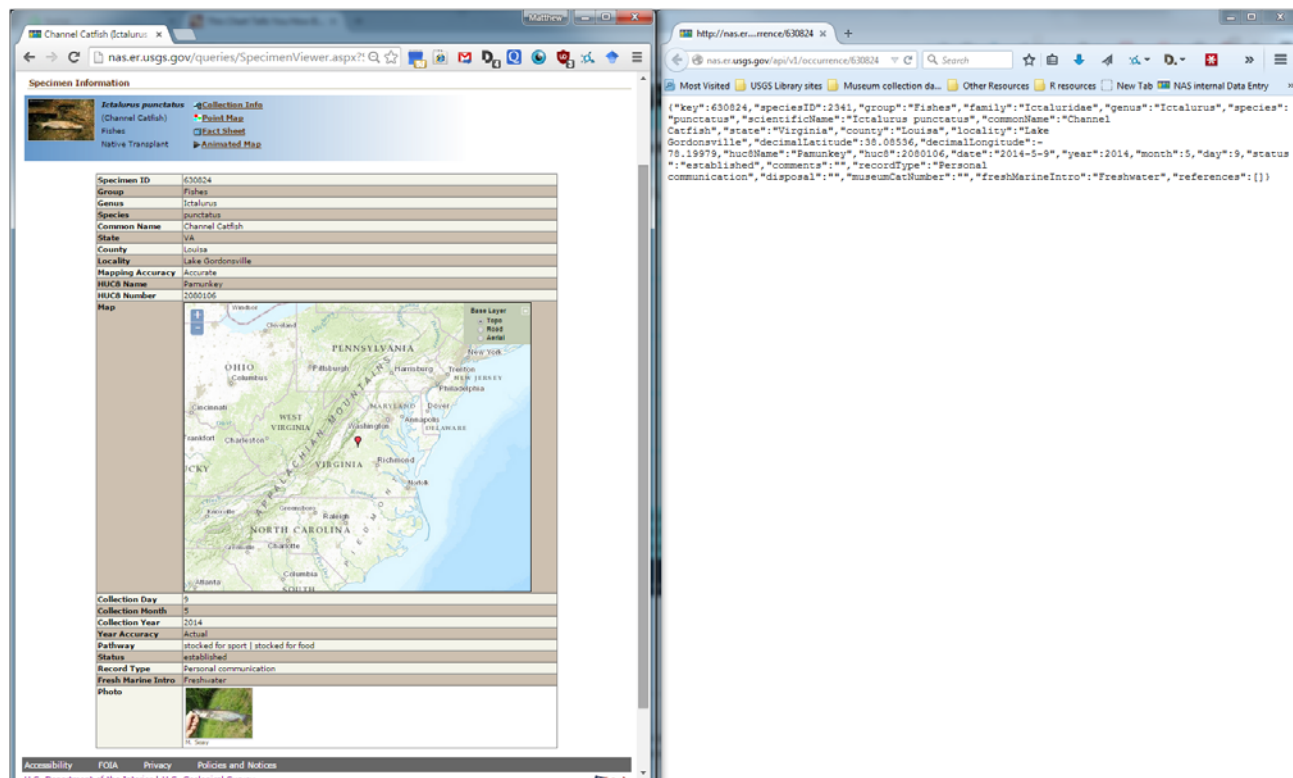


Figure 11. Comparison of the NASWeb Web-accessible specimen record (left) with the textual version that is returned from the NASWeb application programming interface (right).

North American Bat Data Integration

The purpose of this project was to integrate the U.S. Bat Banding Program data (1932–72) and the United States and Canada diagnostic data for WNS with the USGS BPD Project; and provide the bat research community with secure, role-based access to these previously unavailable datasets (U.S. Geological Survey, 2015i). The objectives of this project were to do the following: (1) integrate WNS diagnostic data into the BPD; (2) incorporate the historical bat banding data produced by the Bat Banding Program into the BPD; and (3) develop the APIs and data services required to share these datasets with DOI and USGS enterprise data resources, BISON, and ScienceBase.

Given the differing states of the bat banding and WNS datasets, the project team developed two technical approaches to migrating data and a third for developing Representational State Transfer (REST) services for improved data access. First, the original bat banding data collected from 1932 to 1972 were documented and maintained in 90 drawers of 3x5-inch index cards containing banding information for about 1.5 million individual bats. The file drawers were shipped to the Fort Collins Science Center and scanned into digital format. Data managers transcribed the digital records into the BPD “Observational Data Workbook,” which facilitates batch data uploads to the BPD. As of 2014, about 250,000 individual capture and recaptures have been entered into the BPD.

Our technical approach for the second dataset, the WNS diagnostic data, was simpler because of the source data’s digital format. The project team documented the existing WNS diagnostic data and validated those source data types against the existing BPD data model. Scripts were then used to transfer the existing data to the BPD. The project team developed a “Bat Population Data Project License Agreement” with the USGS Technology Transfer office that protects sensitive bat data (for example, location, disease, and species data).

Lastly, the project team developed a series of REST services that provide structured query options for users to search for BPD data based on their BPD data project access permissions; for example, DOI users can search for their BISON eligible data and export it as a single comma separated value (CSV) file that can be quickly incorporated into the BISON data workbook template for processing.

There are a few benefits of this project. First, by integrating the WNS diagnostic data into the BPD, those data are now available to the WNS Working Groups, other USGS scientists, conservation managers, and the greater research community within a secure, role-based data management environment. Second, when completed, the project will provide 1.5 million indexed and discoverable bat species occurrence records that will be available to bat research community as well as the general public. Third, by creating secure BPD data services, WNS and bat banding data are available to Web and data applications used by bat researchers and policy makers. These benefits directly relate to the CDI SSF by providing applications and Web services that make data and information assets accessible.

Deliverables

The deliverables for the North American Bat Data Integration project were as follows:

- 250,000 records from the historical U.S. Government Bat Banding Program Data (1932–72). The project team will continue entering and importing the remainder of the files in FY 2015–16.
- WNS diagnostic laboratory data. These data have been migrated in their entirety to the BPD; however, because of the sensitive nature of these data, they are not publicly available and can only be accessed by the WNS Data Management Team. The BPD or WNS Teams can provide a live tour of the data to interested parties.

- The BPD Project License Agreement. This agreement is not available online, but by request. The project team plans to provide this document through the BPD by the end of FY 2015.
- BPD Observational Data Workbook and Upload
- BPD Rest Services (Everette, 2015)
- The USGS BPD Project Web page (Ellison, 2015).
- The USGS Bat Population Database product Web page (Ellison and others, 2003)
- USGS Highlight (June 2015): The USGS BPD Project Partners with WNS Laboratories to Improve Disease Tracking and Data Management.
- North American Bat Data Integration Poster presented at the 2014 ESIP Summer Meeting (Ellison and others, 2014).
- Map of WNS spread (fig. 12)

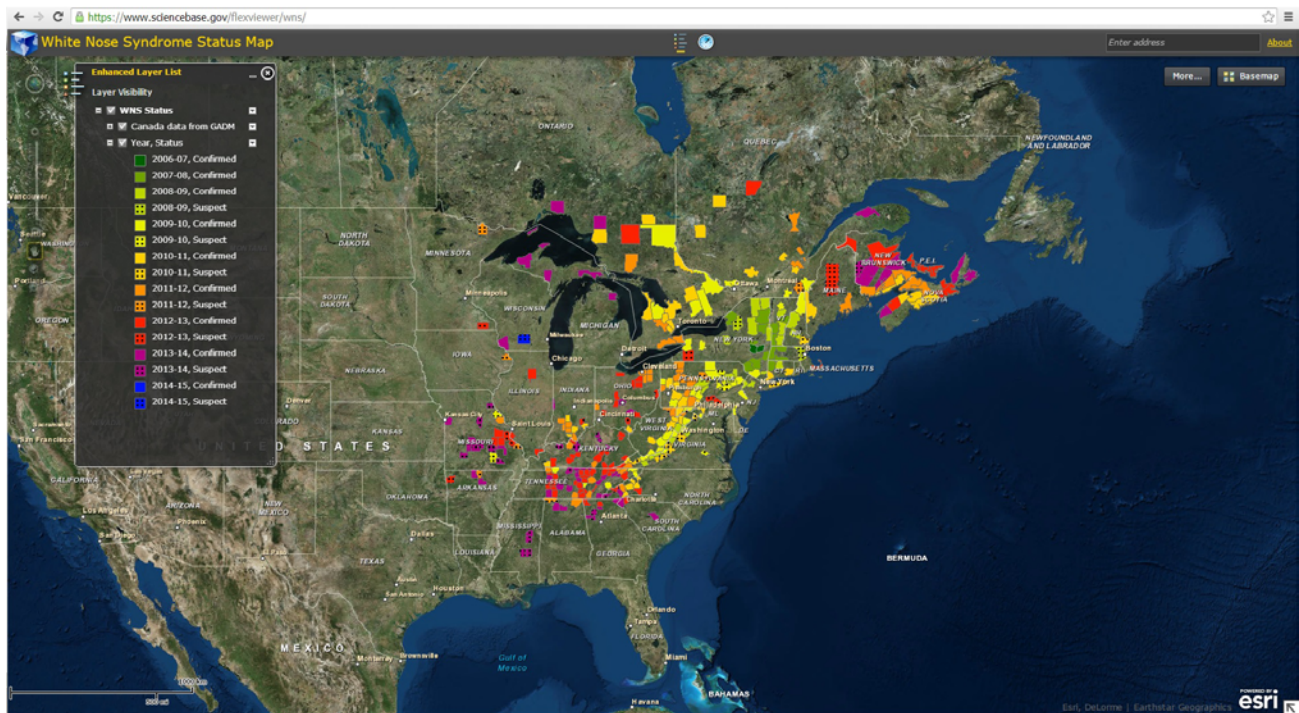


Figure 12. A screen capture of the spread of white-nose syndrome in North America.

Online Merging and Gridding of Topographic and Bathymetric Data Sources

DEMs provide details of the Earth's surface and are used for visualization, physical modeling, and elevation change analysis. Creating DEMs in coastal environments is complicated because of the highly ephemeral nature of the coast and the need to span the land-water interface. This creation of DEMs requires merging multiple bathymetric and topographic datasets that have been collected at different times, using different instrument platforms with varying levels of accuracy, and with variable spatial resolution and coverage. Because coasts can change over short time scales (days to weeks in the case of storms), rapid updates to coastal DEMs are also needed. These challenges and the lack of available tools to create coastal DEMs, lead scientists to sacrifice accuracy by (1) relying on previously

merged, outdated, or over-smoothed DEMs; (2) discarding more recent data that cover only part of the DEM; and (3) using inconsistent methodologies to generate DEMs. Although there are existing projects using sophisticated GIS tools to create DEMs in some specific coastal regions, the process is time consuming and may not meet the needs of individual scientists and resource managers.

The objective of this CDI project was to develop an online, on-demand tool to integrate land and water-based elevation data sources and create DEMs driven by user specifications. To achieve this, the project team developed an ArcGIS Web processing service to format and initialize Python-based geoprocessing tasks (for example, Plant and others, 2002) designed to create on-demand coastal DEMs. The Web processing service is maintained within the USGS ScienceBase data management system and has an associated user interface.

Through the map-based interface, users define a geographic region that identifies the bounds of the desired DEM and a time period of interest. In addition to any user-supplied elevation data, this starts a query for elevation datasets available within Federal science agency data repositories. The geoprocessing service is then triggered to interpolate, merge, and smooth the data sources using the user-defined configuration parameters (resolution, smoothing scales, error thresholds, etc). Uncertainty and error estimates for the DEM are also returned by the geoprocessing service. Upon completion, the information management platform provides access to the final gridded data derivative and saves the configuration parameters for future reference by all users.

The benefits of this online tool include increased access and use of USGS elevation products through easily generated coastal DEMs and reduced overhead costs for coastal researchers and managers who rely on these elevation products. By providing this service, the project team anticipates an increased consistency among coastal elevation models; and by making the geoprocessing codes open source, they can continue to be enhanced by the research community. Also, as a result of this project, the ScienceBase data management system has been improved and is now capable of housing Web processing items, and ingesting and storing new file types.

Deliverables

The deliverables for the Online Merging and Gridding of Topographic and Bathymetric Data Sources project were as follows:

- The framework (fig. 13) was presented at the Federation for Earth Science Information Partners 2014 summer meeting (Frisco, Colo.).
- The project PI was invited to present the project status at the American Geophysical Union 2014 fall meeting (San Francisco, California).
- The Web processing service will be publicly available on ScienceBase. The ability to upload and process input data and store and share processed DEMs is anticipated by August 31, 2015.
- Geoprocessing code dissemination on GitHub is anticipated by May 31, 2015.

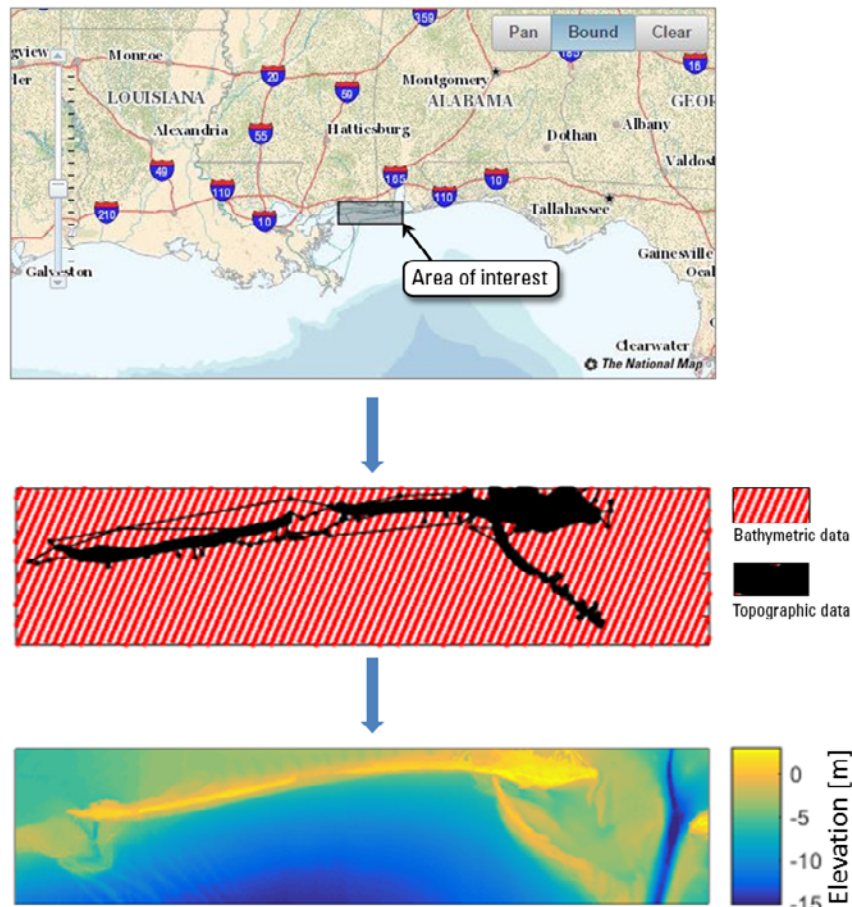


Figure 13. Framework for online, on-demand tool to create coastal digital elevation models. Users define an area of interest using the example Web interface (top panel). Based on a query of elevation resources, relevant datasets in the selected region are identified (middle panel) where, in this example, bathymetric data are shown in red and locations of topographic data in black. Finally, the geoprocessing service merges, interpolates, and smooths the data into a digital elevation model using the user-defined requirements (bottom panel).

Portable ISO 19115 Open Source Developer's Toolkit

In last few years, the International Organization for Standardization (ISO) 19115 family of metadata standards has become the predominantly accepted worldwide standard for sharing information about the availability and usability of scientific datasets among researchers. U.S. interests in the ISO standard have also been growing as global-scale science demands participation with the broader international community. Yet, because of the complexity and rigor of the ISO metadata standards, adoption has been slow. In addition, support for data content in current implementations of ISO 19115 has been minimal.

In 2009, the Alaska Data Integration working group members (ADIWG) mobilized to jointly address the common data integration efforts. Beginning in 2012, ADIWG started to focus on difficulties associated with generating and exchanging ISO 19115–2 metadata among its partners, which vary in size from small nongovernmental organization (NGO) research groups to State offices, universities, and

large Federal bureaus. This differential in size and technical capabilities among its membership brings with it the predictable diversity in metadata requirements and ability to provide the necessary technical support to its researchers.

The ISO 19115 metadata standard is not a single standard, but a family of ISO and Geography Markup Language (GML) standards. Understanding this amalgamation of standards to a degree sufficient to generate a valid metadata record requires an investment of time well beyond what should be reasonably expected of any PI or end user. These issues are not specific to ADIWG and its members; they are shared by many organizations striving to transition to the ISO metadata standard.

After much discussion, the ADIWG membership agreed to coauthor an ISO software toolkit (fig. 14) with an architecture that has the flexibility to achieve the following goals:

1. Isolate the complexity, rigor, formatting, and terminology of the ISO standards from the PI and (or) data steward. Users should not be required to know anything about the ISO standards.
2. Provide clear developer and user documentation.
3. Design a layered architecture so developers will have access to each layer. Developers should be able to enter the toolbox at whatever layer best fits their needs; for example, download the code library to do some custom programming or post JSON metadata to a hosted JSON to ISO translator.
4. Design the architecture to accommodate incremental development cycles. Add features without disrupting previous implementations.
5. Implement all code and services as open-source software projects using a GitHub repository.
6. Use simple JSON for storing and transferring user metadata.
7. Provide flexibility to write to multiple metadata standards including ISO 19115–2:2009, 19115–1:2014, 19110, and possibly Federal Geographic Data Committee (FGDC).
8. Provide an online, interactive end-user metadata editor written in JavaScript to gather and edit a user’s metadata and format it in the “mdJson” format.
9. Host a publicly available version of the online metadata editor for end users to enter and edit metadata, and then request a valid ISO standard metadata record to be returned.
10. Use code that is written in a popular computer language, platform independent, royalty free, and available through an open-source repository to encourage participation from and benefit to the widest possible user base.

To be clear, these tools are for the generation of metadata only. This project is not intended to provide services as a metadata clearinghouse, metadata storage, or to search and modify capabilities. These abilities are deferred to established services such as DataONE and ScienceBase.

Deliverables

The deliverables for the Portable ISO 19115 Open Source Developer’s Toolkit project were as follows:

- “mdTranslator” version 0.11.2 was developed, which supports a reader for the “mdJson” schema and writers for ISO 19115–2 and ISO 19110 metadata.
- “mdTranslator” is publicly hosted with API documentation and a demo.
- A command-line interface (CLI) was added to “mdTranslator” to make the translator accessible from command lines and in particular application servers such as PHP.
- “mdJson-schema” version 0.9.4 definition was peer reviewed by the USGS, ADIWG, and Ted Habermann.

- The “mdJson-schema” definition was extended to include data dictionary information for ISO 19110.
- “mdJson-schema-viewer” version 0.9.4 was developed as an online reference for the “mdJson-schema” structure, definitions, mapping to ISO and FGDC, that also validates “mdJson” files and submits “mdJson” to the “mdTranslator.”
- “mdCodes” version 0.4.0 was developed to deliver standard and extended ISO codelists and custom codelists to “mdEditor” as JSONp objects.
- “mdCodes” was developed to support creation of an ISO CT_CodelistCatalogue record for use in ISO metadata records
- “mdEditor” requirements definitions were written.
- “mdTranslator,” “mdJson-schema,” and “mdCodes” were packaged as Ruby gems.
- “mdBook” version 0.1.0 was written to document all ISO toolkit tools in a GitBook format (Alaska Data Integration Working Group, 2015a).
- Code libraries for the 2013 prototype were upgraded to current versions of Ruby (2.1.0) and Rails (4.1.1).

The “mdCodes,” “mdBook,” and “mdTranslator” writer for ISO 19110 were developed in addition to the promised deliverables for 2014. The “mdJson-schema-viewer” was developed as part of a 2014 extension from CDI. All the products described above can be found on the mdTranslator Web site (Alaska Data Integration Working Group, n.d.)

All code and documentation for the mdTranslator project are made available through public GitHub repositories (Alaska Data Integration Working Group, 2015b). The GitHub directory holds 11 repositories that logically partition the coding projects and their documentation files. Although the repositories are open to the public, only the project administrators (Stan Smith and Josh Bradley) can publish changes to the master branch.

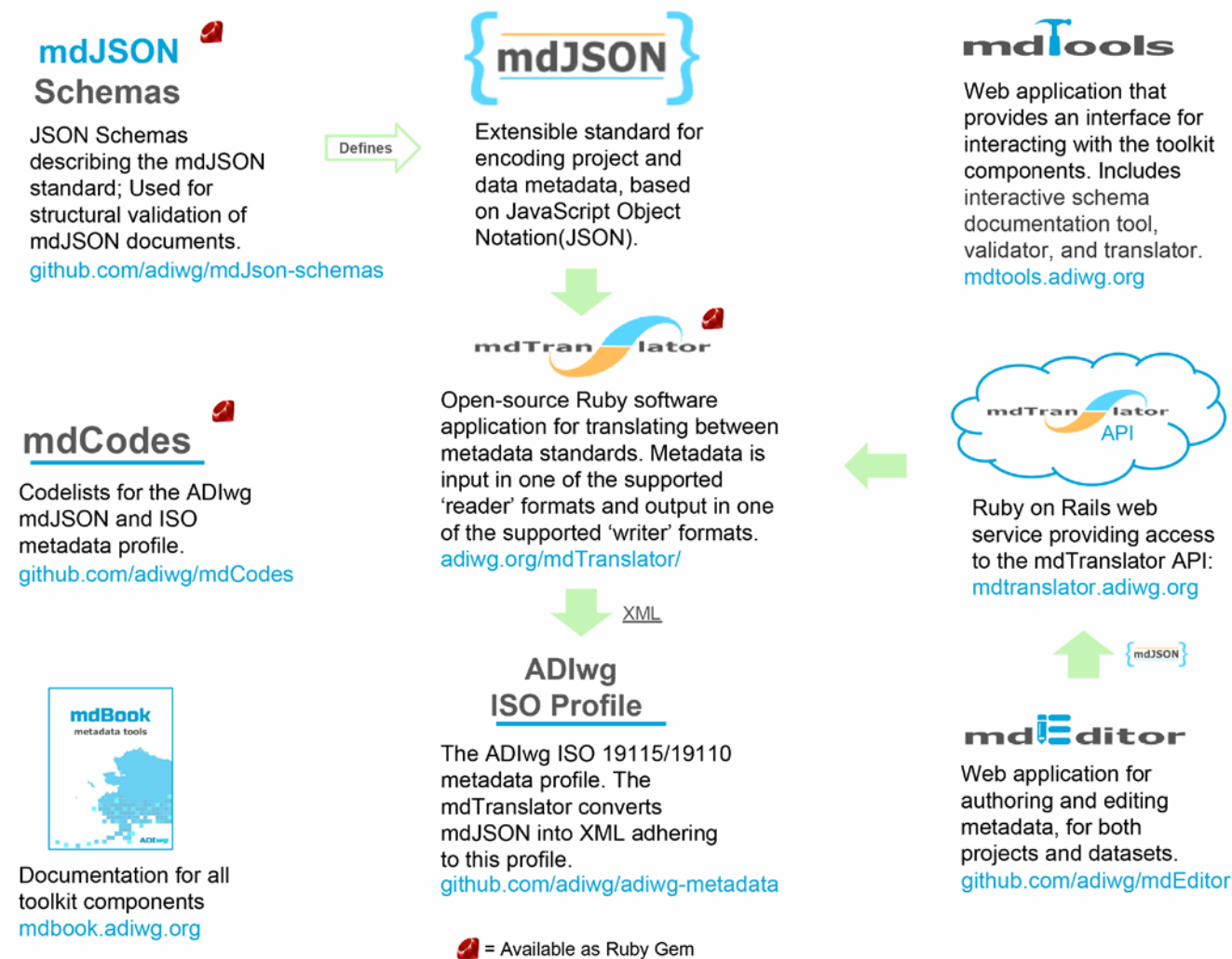


Figure 14. This figure briefly describes each component of the Portable ISO 19115 Open Source Developer's Toolkit and illustrates some possible workflows for generating ISO 19115 metadata using the toolkit.

Summarization of National NEXt Generation RADar Data for Use in Biological Applications

The U.S. network of 160 weather radars known as NEXt generation RADar (NEXRAD) is one of the largest and most comprehensive terrestrial sensor networks in the world. As of 2014, the National Climatic Data Center (NCDC) has archived about two petabytes of data from this system. Although designed for meteorological applications, these radars readily detect the movements of birds, bats, and insects (fig. 15). Many of these movements are continental in scope, spanning the entire range of the network. It is unclear whether biological or meteorological data comprise the bulk of the archive. Regardless, the biological portion is sufficiently large that it likely represents one of the largest biological data archives in the world (perhaps second only to satellite-based remote sensing data).

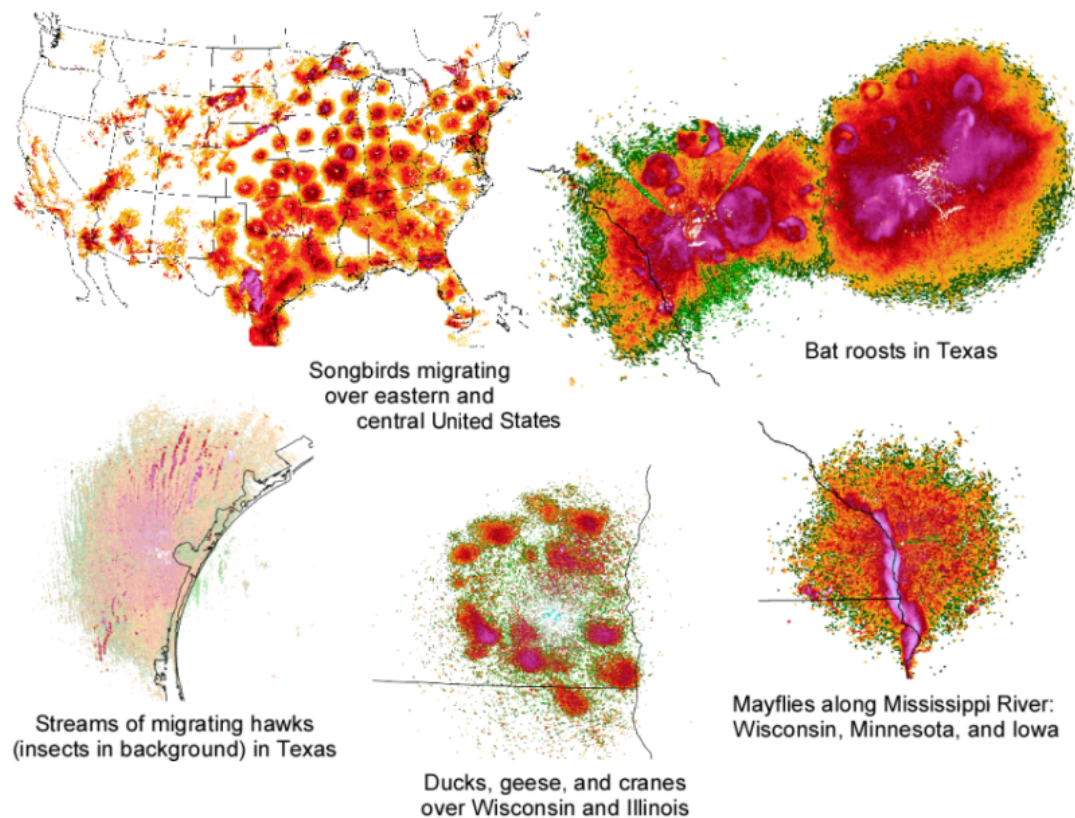


Figure 15. Examples of biological targets detected by NEXRAD.

Ongoing research continues to demonstrate that this radar network shows enormous potential to inform on a wide range of applications including wildlife biology, conservation, and management; weather-related modeling and forecasting; energy development; human safety; and public education (Diehl, 2013). Harnessing this potential of NEXRAD data for biological applications depends on empowering prospective users in gaining access to that subset of the NCDC archive in a usable form. It is a goal of the biological radar research community to advance software and database capability to make these data more accessible to the wider scientific community. To that end, several technological difficulties must still be overcome before the full potential of the NEXRAD data archive is fully realized. Specifically, NEXRAD's potential as biological instrumentation is far from realized owing to the size of the data archive, the challenges faced by prospective users in distinguishing biological from nonbiological data, and the absence of a mechanism for making those data available in a more accessible form. Focusing on this last point (fig. 16, step 4), this project pursued a proof-of-concept demonstrating how biological weather radar data can be postprocessed into value-added biological metrics that are accessible to the public and can address large spatial and temporal scale biological questions. The goal of the project was to show a series of data processing steps toward making derived radar-based biological metrics available to the public for scientific purposes.

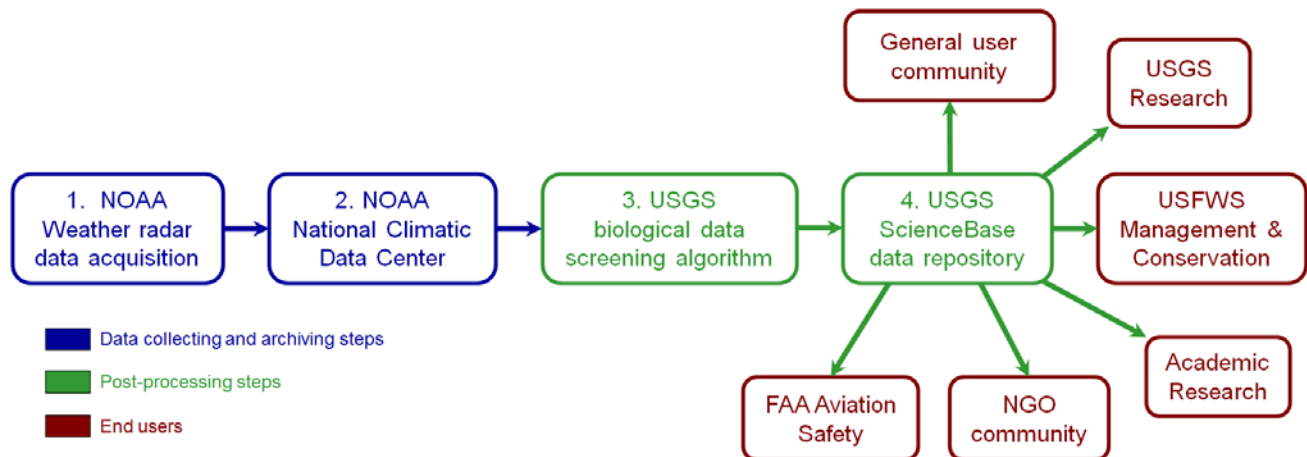


Figure 16. Steps in the process of enabling the biological application of weather radar from data collection and archiving, to postprocessing, to end user.

Deliverables

The deliverables for the Summarization of National NEXRAD Data for Use in Biological Applications project were as follows:

- Proof-of-concept using temporal NEXRAD data to derive biological metrics. The temporal dataset focused on documenting the phenomenon of reverse migration—where birds migrate in a direction that is not seasonally appropriate—over several years. A total of 6 years (more than 350 days) of fall data were retrieved from the Mobile, Alabama, radar station. Fortuitously, data on bird migration from this radar were well positioned geographically to coincide with six corresponding years of existing radio telemetry data on migrating birds that effectively served as ground truth for radar observations of reverse migration. The technologies proved highly complementary. Although radio tracking data show that occasional individual birds engage in reverse migration (Smolinsky and others, 2013), weather radar data show that the phenomenon is much more widespread, with thousands of birds engaging in this activity across an area larger in extent than could be practically covered by other monitoring methods.
- Proof-of-concept using spatial NEXRAD data to derive biological metrics. The large spatial dataset focused on how migrating birds respond to tropical storms in the Gulf of Mexico. Regional/continental-scale snapshots of data were retrieved from more than 20 radars where widespread bird migration was present during the passage of tropical storms of varying intensity (for example, fig. 17). Conventional wisdom considers such storm activity as one of many challenges birds must face during migratory passage. This project showed, however, that tropical storm activity also promotes the movements of birds passing to the west of storm centers where winds are strong and favorable for fall migration. This project also explored how this benefit varies with the intensity of these storms and in relation to typical migratory conditions.
- As a result of this project, multiple metrics on the biology of aero fauna from hundreds of so-called radar volume scans will be uploaded to ScienceBase where they will be publicly available for download, increasing data discovery and integration of NEXRAD data to broader scientific communities within and outside of the USGS. The project advanced methods for the processing,

analysis, preservation, and publishing/sharing of this valuable, but underused data source; and did so in the context of addressing biological questions that will themselves be published in the primary literature.

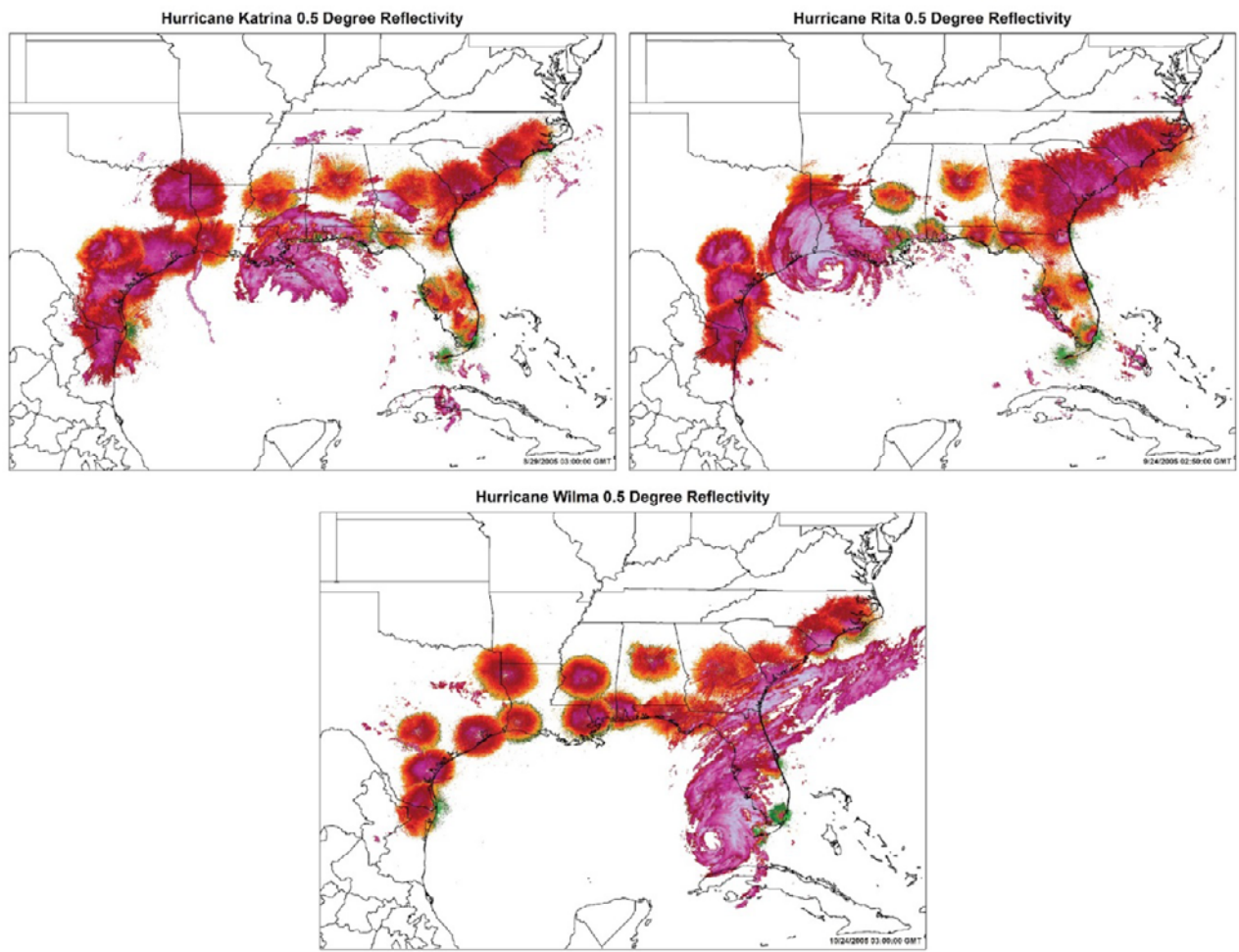


Figure 17. Summarized NEXRAD data from 21 radar stations showing the strength of radar echoes associated with bird migration (circular patterns) in the presence of three strong tropical storms that made landfall during the 2005 hurricane seasons, Katrina (upper left), Rita (upper right), and Wilma (bottom).

Use of Controlled Vocabularies in USGS Information Applications

Large online data catalogs use controlled vocabularies to categorize datasets in ways that allow end users to sort and select data matching their needs. Although the Semantic Web community has developed methods of expressing controlled vocabularies, the community lacks firm standards specifying how Web services should provide vocabularies to end users and applications. The eventual goal of this project is to build functional services so the USGS Thesaurus, Biocomplexity Thesaurus, and other widely used controlled vocabularies will be available to the English-speaking scientific community especially within the USGS where they can be used to improve metadata quality and data

discovery. During FY 2014, use cases were developed and analyzed to understand how vocabulary services can work effectively with existing applications and processes to achieve the advantages of using controlled vocabularies while making the services easier to use for USGS scientists and catalog users. The products of this work will be (1) a white paper detailing the requirements for USGS vocabulary services, (2) a prototype vocabulary server and modifications to existing tools to test the prototype, and (3) a publication presenting a plan for providing USGS authoritative controlled vocabularies through standards-based Web services and modifying three important USGS tools to make use of these vocabularies: the USGS Science Data Catalog, the Core Science Analytics, Synthesis and Libraries (CSAS&L) Online Metadata Editor (OME), and the Portable ISO 19115–2 Metadata Toolkit (see “Portable ISO 19115 Open Source Developer’s Toolkit” above).

Deliverables

The deliverables for the Use of Controlled Vocabularies in USGS Information Applications project were as follows:

- The project was presented at the CDI Meeting on April 9, 2014, and CDI members were asked to suggest “user stories” for additional use cases that should be developed to discover requirements for vocabulary services.
- The team was selected in May and consisted of Alan Allwardt, Janice Gordon, Drew Ignizio, Fran Lightsom, Peter Schweitzer, and Lisa Zolly.
- The Use Cases Face-to-Face Meeting was held in Denver, July 7–11, 2014. In discussing the user stories, the panel chose to develop three use cases different from those originally proposed in order to explore the complete set of functional requirements. The first use case, “Assign keywords to a metadata record using one or more controlled vocabularies,” develops the functional requirements for a vocabulary server interacting with a metadata creation tool. The second use case, “A catalog search interface uses vocabulary services to help users find data,” develops functional requirements for the same vocabulary server, but this time interacting with a catalog user interface. The third use case, “Create specialized indexes to enhance the searchability of metadata,” develops the requirements if the vocabulary server is used by a catalog system to develop an internal table that cleans up and cross-references keywords that are in metadata records.
- By August 15, 2014, documentation was produced for three written use cases including narratives summarizing how “actors” interact with the system to achieve their goals; activity diagrams illustrating the normal flow of events, as well as alternate flows; information models; a diagram summarizing the relationship among the use cases (fig. 3); and a preliminary list of services that would be required to fulfill the use cases.

Summary

Fiscal year (FY) 2014 was another great year for the Community for Data Integration (CDI). The CDI has remained committed to enhancing U.S. Geological Survey (USGS) data and information by collaborating internally and externally, leveraging existing capabilities, sharing expertise, and developing useful products and tools. The CDI Monthly Forums continue to offer members the opportunity to have an open dialogue about data integration challenges, available tools and workflows for data integration, and internal and external data policies and activities. The 2014 CDI Workshop provided community members with the opportunity to meet in person to share information about project

work and to receive feedback. The CDI members who were unable to attend in person had the opportunity to participate remotely. The 2014 CDI Workshop at the ESIP meeting also allowed the CDI to further establish its partnership with the ESIP Federation and many external partners who belong to that group.

In FY 2014, the CDI had five active working groups: Citizen Science Working Group, Connected Devices Working Group, Data Management Working Group, Semantic Web Working Group, and Technology Stack Working Group. These CDI working groups have made many accomplishments this year from the development of new resources to completion of studies. The working groups also coordinated presentations for community members with more focused interests in specific tools, best practices, and policies.

The CDI held its second annual formal RFP process in FY 2014. The CDI incorporated community-wide input into the project selection process for FY 2014. This process has proven to be a successful means to engage the whole community and encourage groups to develop innovative approaches and products to solving our biggest data challenges. The FY 2014 projects represented many elements of the SSF including data, information, communities of practice, applications, Web services, semantics, acquisition, processing, and analysis. Five out of the nine projects developed Web services as one of their deliverables. Additionally, five projects developed new or improved existing applications to better document, integrate, expose, and improve access to data and information. In FY 2014, the CDI Review also identified three proposals that the potential to become much wider reaching and broader in scope, if provided with additional time and funding. The CDI Review Panel asked the applicants of these proposals to adjust their proposals to reflect these larger perspectives.

Overall, the CDI has had a highly successful year and continues to bring researchers, data managers, and information technologists together to sustain the CDI Science Support Framework in support of USGS Earth science research. Membership is voluntary and open to USGS employees and other individuals and organizations willing to contribute to the community. Interested parties are invited to visit the CDI Web site at <http://www.usgs.gov/cdi/>. Contact cdi@usgs.gov for more information about joining the CDI.

References Cited

- Alaska Data Integration Working Group, 2015a, mdBook—Documentation for ADIWG open source metadata tools: Alaska Data Integration Working Group, accessed August 27, 2015, at <https://www.gitbook.com/book/adiwg/mdbook/details>.
- Alaska Data Integration Working Group, 2015b, Alaska Data Integration Working Group—Working on methods to efficiently integrate and share data: Alaska Data Integration Working Group, GitHub repository, accessed August 27, 2015, at <https://github.com/adiwg>.
- Alaska Data Integration Working Group, [n.d.], Welcome to the mdTranslator: Alaska Data Integration Working Group, accessed August 27, 2015, at <http://mdtranslator.adiwg.org/>.
- Atzori, Luigi, Iera, Antonio, and Morabito, Giacomo, 2010, The internet of things—A survey: Computer Networks, v. 54, no. 15, p. 2787–2805, accessed August 27, 2015, at <http://dx.doi.org/10.1016/j.comnet.2010.05.010>.
- Bowser, Anne, and Shanley, Lea, 2013, New visions in citizen science: Washington, The Woodrow Wilson Center, 53 p., accessed August 27, 2015, at <https://www.wilsoncenter.org/sites/default/files/NewVisionsInCitizenScience.pdf>.
- Diehl, R.H., 2013, Realizing the biological potential of weather radar: U.S. Geological Survey, Northern Rocky Mountain Science Center, accessed August 27, 2015, at http://www.nrmssc.usgs.gov/staff/diehl/research/remote_sensing/weather_radar.

- Ellison, L.E., 2015, Bat population data project: U.S. Geological Survey, Fort Collins Science Center, Research Project RB00CN7.2.0, accessed August 27, 2015, at <https://www.fort.usgs.gov/science-tasks/2217>.
- Ellison, L.E., Everette, A.L., Ballmann, A.E., Peurach, S.C., and Coleman, J., 2014, North American bat data integration [abs.], in Federation of Earth Science Information Partners meeting, Frisco, Colorado, 15 July 2014: Federation of Earth Science Information Partners, p. 1., accessed August 27, 2015, at <https://www.fort.usgs.gov/products/23762>.
- Ellison, L.E., O'Shea, T.J., Bogan, M.A., Everette, A.L., and Schneider, D.M., 2003, USGS bat population database for the United States and trust territories (BPD) v.1.0: U.S. Geological Survey, Fort Collins Science Center, accessed August 27, 2015, at <https://www.fort.usgs.gov/products/23233>.
- Everette, A.L., 2015, 2.2—Automated data management: U.S. Geological Survey, accessed August 27, 2015, at <https://my.usgs.gov/confluence/pages/viewpage.action?title=2.2:+Automated+Data+Management&spaceKey=fortproj>.
- 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 August 27, 2015, at <http://dx.doi.org/10.3133/ofr20131265>.
- Lightsom, F.L., 2014, Vocabulary service use cases—Products of the 2014 SWWG project: U.S. Geological Survey, accessed August 27, 2015, at <https://my.usgs.gov/confluence/display/cdi/Vocabulary+Service+Use+Cases>.
- Plant, N.G., Holland, K.T., and Puleo, J.A., 2002, Analysis of the scale of errors in nearshore bathymetric data: *Marine Geology*, v. 191, no. 1–2, p. 71–86. [Also available at [http://dx.doi.org/10.1016/S0025-3227\(02\)00497-8](http://dx.doi.org/10.1016/S0025-3227(02)00497-8).]
- U.S. Environmental Protection Agency, 2015, Federal community of practice for crowdsourcing and citizen science: U.S. Environmental Protection Agency, accessed August 27, 2015, at <http://www2.epa.gov/innovation/federal-community-practice-crowdsourcing-and-citizen-science>.
- U.S. Geological Survey, 2013, US Geological Survey (USGS) Community for Data Integration (CDI) request for proposals (RFP)—For fiscal year 2014: U.S. Geological Survey, accessed August 27, 2015, at https://my.usgs.gov/confluence/download/attachments/423788559/CDI%20FY14%20Request%20for%20Proposals%20FINAL_Updated.pdf?version=1&modificationDate=1382368768967&api=v2.
- U.S. Geological Survey, 2014a, Distinctions between new research or interpretive information products and previously published or noninterpretive information products: U.S. Geological Survey, Fundamental Science Practices (FSP), accessed August 27, 2015, at http://www.usgs.gov/fsp/interpretive_definitions_and_examples.asp.
- U.S. Geological Survey, 2014b, Distinguishing noninterpretive, interpretive and new interpretive USGS information products: U.S. Geological Survey, accessed August 27, 2015, at http://www.usgs.gov/fsp/interpretive_decision_workflow.pdf.
- U.S. Geological Survey, 2014c, Frequently asked questions about digital object identifiers (DOIs)—For USGS researchers: U.S. Geological Survey, 6 p., accessed August 27, 2015, at http://www.usgs.gov/datamanagement/dama/internal_only/digital-object-identifiers-for-usgs-faq.pdf.
- U.S. Geological Survey, 2015a, U.S. Geological Survey Community for Data Integration (CDI) science support framework (SSF): U.S. Geological Survey, 3 p., accessed August 27, 2015, at <http://www.usgs.gov/cdi/cdi-ssf/cdi-ssf-components.pdf>.

- U.S. Geological Survey, 2015b, USGS data management—Supporting and enabling USGS data management: U.S. Geological Survey, accessed August 27, 2015, at <http://www.usgs.gov/datamanagement/>.
- U.S. Geological Survey, 2015c, USGS data management—USGS science data exit survey form: U.S. Geological Survey, accessed August 27, 2015, at <http://www.usgs.gov/datamanagement/preserve/exitsurvey.php>.
- U.S. Geological Survey, 2015d, U.S. Geological Survey Manual—Instructional memoranda: U.S. Geological Survey, accessed August 27, 2015, at <http://www.usgs.gov/usgs-manual/95imlist.html>.
- U.S. Geological Survey, 2015e, USGS data management—Data release: U.S. Geological Survey, accessed August 27, 2015, at <http://www.usgs.gov/datamanagement/share/datarelease.php>.
- U.S. Geological Survey, 2015f, Community for Data Integration (CDI)—Awarded 2014 projects: U.S. Geological Survey, accessed August 27, 2015, at <http://www.usgs.gov/cdi/year/2014.html>.
- U.S. Geological Survey, 2015g, NAS—Nonindigenous aquatic species: U.S. Geological Survey, accessed August 27, 2015, at <http://nas.er.usgs.gov/>.
- U.S. Geological Survey, 2015h, NAS Web services API: U.S. Geological Survey, accessed August 27, 2015, at <http://nas.er.usgs.gov/api/documentation.aspx>.
- U.S. Geological Survey, 2015i, The USGS bat population data project: U.S. Geological Survey, accessed August 27, 2015, at <https://my.usgs.gov/bpd/>.
- Wilson Center's Science and Technology Innovation Program, 2014, An exploratory study on barriers: Commons Lab—Mobilizing Public Participation, accessed August 27, 2015, at <http://wilsoncommonslib.org/2014/09/07/an-exploratory-study-on-barriers/>.