CDI SSF Category:
Computational Tools and Services

Mining the USGS Data Landscape

Applicants/Principal Investigator(s):
Lance Everette, USGS Fort Collins Science Center, 2150 Centre Ave, Fort Collins, CO 80526
Ph (970) 226-9225 Email everettel@usgs.gov
Susan Skagen, USGS Fort Collins Science Center, 2150 Centre Ave, Fort Collins, CO 80526
Ph (970) 226-9366 Email skagens@usgs.gov

Abstract:
The scientific legacy of the USGS is the data, and the scientific knowledge derived from it gathered over 130 years of research. However, it is widely assumed, and in some cases known, that high quality data, particularly legacy data critical for large time-scale analyses such as climate change and habitat change, is hidden away in case files, file cabinets, and hard drives housed in USGS science centers and field stations (both hereafter "science centers"). Many USGS science centers, such as the Fort Collins Science Center, have long, established research histories, are known repositories of data sets, and conduct periodic "file room cleanout" days that establish and enforce some minimal data lifecycle management and maintains a cursory inventory of maintainable data – data that is of high enough interest/impact that they should be maintained at a minimum readable format for future access and use. But science centers currently lack a clear understanding of data lifecycle management best practices and simple inventory tools to manage their data through its lifecycle. We propose testing the CDI lifecycle framework by applying it to a handful of known data, and documenting the considerations and requirements of effectively applying the CDI data lifecycle framework. Further, we propose creating a simple "USGS Data Mine" tool that enables science centers to conduct and maintain their data inventories, while contributing to and assisting with the growing greater USGS data landscape.

Total funding amount requested: $80,500
Total in-kind funding: $25,800

Specific Datasets Exposed/Augmented/Distributed:

- Southeastern Arizona riparian area bird and habitat data (1989-1993)
- TX, KS, OK, SD, ND wetland and shorebird data (1989-2011)
- Eastern Colorado prairie bird and habitat data (1997-2012)

Geographic/geologic/ecosystem/habitat/taxonomic/other context:
Western U.S./prairie potholes and wetland ecosystems/avian breeding, occurrence and abundance data.

Type of Product(s) Generated:

- USGS Open-file Report documenting the USGS Data Mine project, user training and help for the Data Mine inventory, with case studies on FORT's inventory process and CDI's testing and validation of the lifecycle management framework.
- USGS Highlights at the completion of each phase (x3)
- Fact Sheet and web-based science feature on the USGS Data Mine Project

- Open-source Data Mine inventory code and database

Summary

Introduction and Background:

USGS is one of the largest science and research organizations in the world and for more than a century its scientists have been collecting vast arrays of data. In its original time and context USGS's data has been analyzed and reported in thousands of products and publications, making USGS one of the most productive scientific organizations in the world (Khan and Ho 2012). However, data collection and analysis are only parts of the data management lifecycle. Other lifecycle phases (planning, description, preservation and sharing/distribution) are critical and designed to identify the true long-term costs of managing science data and ensure long-term data discovery, access and use. More and more, complex natural resource issues require complex, long-term data sets such as those produced by career USGS scientists. Properly applying a data management lifecycle framework allows future scientists to effectively discover those data sets, evaluate their potential utility and cost to implement, access and incorporate it for use in new, novel analyses.

USGS science centers and research stations are known in many cases to have become rich deposits of data, particularly legacy data, with decades-long histories of research in various disciplines. Over years, these USGS "data mines" accumulate volumes of data often stored neatly in science center case files, file rooms and offices but effectively lost to any additional use. There currently is no way of estimating the true total USGS "data wealth" given these hidden data stores, but it isn't difficult to imagine the potential data produced during its 130+ years of research across dozens of science centers and research stations, all potentially data rich and ideally all needing to be inventoried.

Some USGS science centers already mandate minimum data preservation requirements and conduct periodic case file/archive data inventories and reviews. However, beyond these minimum requirements, science centers and their research staff struggle with what they can do to best preserve and distribute even their most significant data. Understanding the requirements, options and potential costs of properly managing data throughout its lifecycle would help them best prioritize and manage their data inventory, benefiting the scientist, science center, and USGS as a whole.

The USGS Center for Data Integration's (CDI) Data Management Working Group recently developed a USGS data management lifecycle framework specifically designed to assist scientists with planning, describing, preserving, and sharing their data (CDI 2012) through its lifecycle. The Fort Collins Science Center is conducting legacy data inventory February 11-15, 2013 and is interested in using that exercise to test, validate, and document their application of the CDI lifecycle framework to inventory and prioritize their Center's diverse datasets. In addition, to estimate time and resource costs for data mine inventory items, we propose developing estimated costs by applying the framework to several significant, existing USGS datasets. Based on the results of these two objectives, we propose creating a simple "USGS Data Mine" tool that enables science centers to conduct and maintain their data inventories, while contributing to the greater USGS inventory. A short Open-file Report will be produced at the conclusion of the project that includes the Data Mine project's overview, inventory application training and documentation, and case studies describing the FORT and CDI experiences applying the CDI data management lifecycle framework.

References

Center for Data Integration Data Management Working Group. 2012. A USGS Website to Support and Enable Better Science Data Management (poster). Available online:

https://my.usgs.gov/confluence/display/cdi/Posters+-+2012.
Khan, MA, and Y-S Ho. 2012. Top-cited articles in environmental sciences: Merits and demerits of citation analysis. Science of the Total Environment 431:122-127.
O'Donnell, M.S. and D.A. Ignizio. 2012. Bioclimatic predictors for supporting ecological applications in the conterminous United States: U.S. Geological Survey Data Series 691. Available online: http://www.fort.usgs.gov/Products/Publications/pub_abstract.asp?PubID=23502
O'Shea, TJ. 2006. History of the Fort Collins Science Center U.S. Geological Survey: U.S. Geological Survey Open-File Report 2006-1336. 27 p.

CDI SSF Category:

- Computational Tools and Services

Project Title:

- Mining the USGS Data Landscape

Contacts:

- Lance Everette, USGS Fort Collins Science Center, 2150 Centre Ave, Bldg C, Fort Collins, CO 80526
- Susan Skagen, USGS Fort Collins Science Center, 2150 Centre Ave, Bldg C, Fort Collins, CO 80526

Collaborating Organizations:

- USGS Fort Collins Science Center

Scope

There are 4 primary objectives to this project, each building on the other, each with a distinct narrow scope:

1. Validate and document the application of the CDI Data Management Lifecycle framework to known legacy data sets.
   The scope of work for this objective is limited to evaluating and processing Dr. Skagen's 3 avian ecology datasets and would be cost-shared between CDI and FORT.
2. Inventory, prioritize and estimate the cost of integrating a USGS data mine.
   The scope of work for this objective is limited to a single USGS science center (FORT), and will be completed entirely through FORT's cost share funding.
3. Develop a USGS Data Mine Inventory web application
   The scope of work for this objective is limited to a simple web application that consists of basic data management functionality (create, review, update, delete) for managing USGS legacy data through the data management lifecycle.
4. Promote the USGS Data Mine project and encourage USGS science centers to conduct and add to the USGS data inventory.
   The scope of work for this objective is limited to the production of USGS highlights and email campaigns at the completion of each objective. In addition, a web-based science feature will be produced at the conclusion of objective 3. At the conclusion of the project an Open-file Report and USGS Publication Brief will be produced.

This effort will:

- Deliver an immediate benefit to solve an existing data integration challenge;
- Demonstrate a methodology and/or solutions architecture that can be repeated/replicated for other data or research projects;
- Create an environment that allows future innovative applications to access USGS data;
- Provide a benefit to scientists;
- Promote standards and best practices for data management; and
- Complement other SSF categories, elements, and CDI projects.

Technical Approach

Each of this project's 3 objectives has a unique set of methods and products described below. All of the technical work will be completed by the FORT's Web Application Development team in Fort Collins, Colorado.

Objective 1: Validate and document the application of the CDI Data Management Lifecycle framework (2 months)

We will apply the data management lifecycle framework to Dr. Susan Skagen's avian ecology data, which spans 20 years of research. With the assistance of Dr. Skagen, the FORT Data Steward will complete the 4 phases of the data management lifecycle framework.

Planning: Our team will gather data attributes and requirements for each of the (e.g., disk requirements, format, archive type). These will become the core of the "Planning" data entry form(s) constructed for the Data Mine inventory application (Objective 3).

Description Phase: Our team will verify that all necessary metadata is created and stored in Sciencebase to ensure long-term data discovery.

Preservation Phase: Our team will ensure that, at a minimum, the following is completed for each data set during the Preservation phase:

- Backups created and (where appropriate/possible) project case files updated (NOTE: data from projects conducted prior to current BASIS+ implementation in 2000 may not have project records to update).
- Data set archived and stored in Sciencebase.
- Disposition documented and verified.
- Persistent data object identifiers obtained and metadata updated.
- Identify additional data repositories to potentially share data with.

Sharing/Distribution: Sharing and distribution will only be tested on a single Skagen data set (to be determined; NOTE: should there still be funds remaining after all other project commitments are met, additional Skagen could be processed and shared). Once selected, we will use the new Sciencebase Expando Facet to model and test Sciencebase custom facet requirements needed for ingesting row-level data. Once the custom facet is created, we will develop a harvester and ingest the row-level data into the custom facet.

Funding: FORT Cost-share and CDI

Product: Case Study of the application of the CDI data management lifecycle framework on 3 legacy data sets.

Data exposed:

- Southeastern Arizona riparian bird and habitat data (1989-1993),
- Texas, Kansas, Oklahoma, South Dakota, North Dakota wetlands and shorebird data (1989-2011),
- Eastern Colorado prairie bird and habitat data (1997-2012)

Objective 2: Inventory, prioritize and estimate the cost of integrating a USGS data mine. (3 months)

Objective 2 is designed to take the lifecycle workflow, cost, and resource requirements documented in Objective 1 and assess the estimated cost and requirements of integrating a science center's entire "data mine" into a USGS Data Mine Inventory. The Fort Collins Science

Center will conduct the initial inventory February 11-15, 2013 using 3 methods of data discovery:

1. Known: the FORT will conduct a survey of current science project investigators to identify science data whose project is complete and the data is no longer being analyzed as part of that work. This survey will be funded by the FORT as part of their scheduled annual science data/metadata review process.
2. Assumed: FORT has 218 metadata records documented. Assuming all of these can be located, these would immediately give FORT a significant inventory of potential.
3. Hidden: The FORT has two known "informal" locations where data has also accumulated over years – particularly older, pre-BASIS+ data (pre-2000). The FORT data steward will go through these file cabinets file-by-file looking for to assess for data management lifecycle.

All data sets identified in this inventory process will be assessed for their data management lifecycle needs (i.e., conduct the data management lifecycle's planning phase). Based on the cost estimates generated from the project's work in Objective 1, we will calculate the estimated total cost of applying the data management lifecycle framework to each data set inventoried, as well as a total estimated cost to apply it to all of the FORT's data. We will also work with the FORT to prioritize which data to integrate and in what order to distribute total cost over time effectively.

Funding: FORT cost-share (NOTE: no CDI funding requested)

Product: Documentation of the methods used and results of the assessment of the FORT as a USGS data mine.

Objective 3: Develop a USGS Data Mine Web application (3 months)

Based on the lifecycle workflows documented in Objective 1 and the data mine assessment methods documented in Objective 2, a simple web application will be developed to enable science centers to conduct legacy data inventories and prioritize data sets to apply the CDI data management lifecycle to. The web application will also act as a USGS legacy data discovery tool that allows users to search the USGS data landscape and access available data.

Planning Phase: scientists create a "data plan" that collects a project's basic data requirements and calculates an initial estimated cost of the data lifecycle requirements.

Description Phase: Data managers will be able to document their data project and create FGDC compliant metadata and store them in Sciencebase

Preservation Phase: scientists will use a checklist of required and optional items (Persistent DOI, Additional suggested repositories, Disposition plan/schedule) and a file uploader to transfer the data set archives to Sciencebase for minimum data access by end-users.

Sharing/Distribution: scientists will use the Sciencebase Expando Facet and Sciencebase services to model and test Sciencebase facet requirements.

Data Discovery: public-facing data discovery features including taxonomic, geospatial and science center indexes, as well as full-text search. If a data set is accessible online, users are provided a link to it; if the data is *not* online, a data request can be submitted to prioritize the data set for CDI framework application.

Products:

- Data Mine Inventory web application, including systems and user documentation and application code, and a complete inventory of FORT's historic science data mine.

All project and user support documentation will be provided using myUSGS-

Confluence. myUSGS-Jira will be used to manage project tasks and provide real-time updates on task-level progress.

Project Experience

All project management and technical development will be completed by the Fort Collins Science Center's Web Application Development Team, a significant contributor to many USGS Enterprise IT projects (e.g., Sciencebase, The John Wesley Powell Center for Analysis and Synthesis, USGS Mobile Applications, myUSGS).

Commitment to Effort

Fort Collins Science Center has already begun implementing the CDI data management lifecycle during their project planning phases. FORT has also recently hired a Data Steward whose primary responsibility is to assist FORT science staff manage their data through the entire data management lifecycle. All of the work described in Objective 2 of this proposal is scheduled to be completed by the FORT February 2013 regardless of the success of this proposal's funding request. FORT also recognizes that they benefit with the long term success of the Data Mine Inventory project and is contributing writing, editing, and data stewardship staff to promote the project and assist with identifying new science center's interested in participating.

Budget

Budget Category	Federal Funding "Requested"	Matching Funds "Proposed"

1. SALARIES (inc. number of hours and hourly rate):

Federal Personnel		
Dr. Susan Skagen, Scientist/Data Owner	$14,000 (Objective 1)	
Lance Everette, Lead Project Manager	$8,800 (Objective 3)	$8,800 (Objectives 1, 2)
FORT Data Steward		$8,800 (Objectives 2, 3)
Contract Personnel		
Communications Lead (80hrs@$40/hr)		$3,200 (Objectives 1-3)
Sciencebase ExpandoFacet Support (80hrs@$45/hr)	$3,600 (Objective 1)	
Lead Developer (320hrs@$70/hr)	$22,400 (Objective 3)	
Total Salaries:	**$48,800**	**$20,800**

2. FRINGE BENEFITS:

Personnel	$0	$0
Contract Personnel	$0	$0
Total Fringe Benefits:	**$0**	**$0**

3. TRAVEL EXPENSES*:

Per Diem	$1,000	
Airfare	$2,500	
Lodging Cost	$1,000	
Vehicle Cost	$400	
Other travel expense(s)	$100 (airport shuttle)	
Total Travel Expenses:	**$5,000**	**$0**

4. OTHER DIRECT COSTS: (itemize)

USGS Enterprise Publishing Costs for 1 OFR/ITR	$5,000	$5,000
Total Other Direct Costs:	**$5,000**	**$5,000**
Total Direct Costs:	***$58,800***	***$25,800***
Indirect Cost (%)	***$21,700***	***$0***
GRAND TOTAL:	***$80,500***	***$25,800***

*NOTE: all statements of work for contract staff/positions listed above are managed by FORT's Web Application Development Team and these estimates are in accordance with their requirements.

Timeline

Deliverable	Estimated Delivery Date
Objective 1 Completion Report, Exposed Skagen data, USGS Highlight	8 weeks from time of award
Objective 2 Completion Report, USGS highlight	12 weeks from time of award
USGS Data Mine Web Application, documentation, application code, USGS highlight, FORT Science Feature	20 weeks from time of award
Open-File Report	24 weeks from time of award