SECTION 1. PROJECT ADMINISTRATIVE INFORMATION

CDI Science Support Framework Elements: Management Policy and Standards

Lead USGS Cost Center: Fort Collins Science Center, Information Science Branch

Title: Facilitating the USGS Scientific Data Management Foundation by integrating the process into current scientific workflow systems.

Main PI: Colin Talbert
USGS Fort Collins Science Center
2150 Centre Ave.
2150 Centre Ave. Bldg. C
talbertc@usgs.gov

Co PI: Drew Ignizio
USGS Core Science Systems
Mail Stop 302
W. 6th Ave Kipling St.
Lakewood, CO 80225
dignizio@usgs.gov

Co PI: Catherine Jarnevich
USGS Fort Collins Science Center
2150 Centre Ave. Bldg. C
Fort Collins, CO 80526
jarnevichc@usgs.gov

Co PI: Jeff Morisette
DOI North Central Climate Science Center
NESB A309 Colorado State University
Fort Collins, CO 80523
morisettej@usgs.gov

Description: Increasing attention is being paid to the importance of proper scientific data management, and implementing processes that ensure products being released are properly documented. New USGS rules are being established to properly document not only publications, but also the related data ***and*** software. This relatively new and recent expansion of documentation requirements for data and software may present a daunting challenge for many USGS scientists whose major focus is their physical science (and less expertise in information science). Our proposed work will use a current study and existing scientific workflow architecture to provide an exemplar that the USGS can point to as it initiates new standards for producing repeatable science. Providing such an exemplar could help lower anxiety and ensure compliance pertaining to new requirements. We propose to leverage existing services, infrastructure and tools from within an existing scientific workflow system used in the USGS for species distribution modeling to provide a working example of how to integrate and facilitate USGS data management and release policies.

SECTION 2. ESTIMATED BUDGET

Budget Category	Federal Funding	Matching funds
1. Personnel (Salaries)		
Colin Talbert (100 hrs total)	0	5,350
Drew Ignizio (40 hrs total)	0	2,560
Catherine Jarnevich (60 hrs total)	0	3,900
Jeff Morisette (40 hrs total)	0	3,685
Other staff (Developer 460 hrs total)	30,000	0
Total Salaries:	**30,000**	**15,495**
2. Travel Expenses		
CDI Workshop	300	
Travel Total:	**300**	
Total Direct Costs	**30,300**	**15,495**
Indirect Costs (%)	4,605 (15.2%)	
Grand Total	**$34,905**	**$15,495**

Recent years have seen a heightened interest in improving scientific rigor by implementing robust data management practices. The aim has been to improve the transparency, repeatability, and efficiency of the scientific process. Being able to accurately interpret, use, repeat, or build off of previous research requires thorough documentation of the process used and the output produced. This documentation must include meticulous detail about the source data that went into analysis, methods used, which version of processing software was used, what parameters were used, what format the output data are in, etc. Capturing and recording these details in sufficient detail can be a challenge given the increasing use of multiple data sources and complex workflows involving various processing steps each with specific parameters. Within the USGS this has been implemented as a series of policies designed to encourage and assist with this goal (http://www.usgs.gov/fsp/policies.asp).

While the institutional will to improve scientific data management and documentation is obviously present, many scientists do not currently have the experience or tools necessary to fully and efficiently work through the process. Tools to assist in various aspects of this process have been rapidly developing (ScienceBase, MetadataWizard, etc) but the data management process still requires some specialized knowledge, and navigating through multiple tools and interfaces, as well as the details of the USGS requirements.

The development and use of scientific workflow systems has the potential to facility many aspects of the data management process. These software systems integrate existing and new data access and processing tools into a unified and user-friendly system that automatically records much, if not all, of the information needed to interpret and recreate the processing steps; that is, the analysis provenance. Besides providing automated recording of provenance, scientific workflow systems have additional benefits such as real-time visualization, easier sharing of methodologies, and an ability to more easily access remote data and processing resources.

One example of such a scientific workflow system being used in the USGS is the Software for Assisted Habitat Modeling (SAHM) package that has been implemented in VisTrails at the Fort Collins Science Center. This software integrates all steps in the widely used process of species distributions modeling. These statistical models correlate observations of where species have occurred with climate or other GIS and remotely sensed data to produce models of where the species could occur and which environmental factors might be driving or limiting the response. The outputs from these models are used to inform questions about invasive species spread, habitat conservation needs, and projected impacts from climate change. SAHM has been used in numerous USGS projects including modeling migratory bats, invasive plants in Alaska, and white bark pine in Yellowstone.

While using a scientific workflow system will automatically collect most of the data required to support a robust scientific data management plan, using it to feed this process still requires significant manual manipulation. For example a scientific workflow file has a record of the data produced but a used would still need to mint a DOI for this output, create a metadata record for it and upload it to a persistent archive. But each of these steps could be readily automated using existing services and tools. For example, a user-friendly interface accessed from within the scientific software used to create the data could access a web-based API to minting a DOI for the product, extract the bulk of the information required for compliant metadata, walk the user through a metadata creation tool such as the MetadataWizard, to finalize the metadata, and automatically upload the product package to persistent archive such as ScienceBase. This process would be relatively painless to the scientist using it, but at the same time provide a higher quality data product by ensuring that documentation was complete and accurate based on the actual provenance of the data produced.

While species distribution modeling is widely use, within USGS and among the ecological community, it is still only one of the many modeling and analysis tools being used by USGS. However, VisTrails has been used for many other types of analysis, including environmental sciences, psychiatry, astronomy, cosmology, high-energy physics, and quantum physics. As such, we believe the lesson learned from this exemplar, using species distribution modeling, but focusing on the VisTrails scientific workflow tools will be fairly generalizable to another types of work done throughout USGS and the knowledge gained through this project will pay dividends to a broader set of general users. This proposal will deliver an example of how to efficiently integrate the USGS Data Management Policies with tools scientists are already comfortable with. Additionally the core tools developed for this project will be coded in a common processing language such as Python or R. This core functionality could then be exposed to other projects fairly easily by wrapping this functionality in a lightweight package.