2011 Proposals

FY 2011 Annual Report (MSWord) Feb 3, 2012

2011 Funding Deliverables [docx]

FY 2011 Project Proposal Suite

> See the "Attach" tab, above, for the draft high-level **CDI FY 2011 Science Plan** synthesizing the ideas, below, as well as a **concept map** summarizing the current proposal suite's foci and relationships. The concept map discusses focus areas that may be generally accomplished within the scope of the "Community of Practice Facilitation" discussion below. A struggle with the proposal process has been the tension between the free floating activities of working groups with rather ill defined end goals and the need to plan funding with inherent goals and outcomes. The Guiding Principles and other new sections below are one method of trying to balance in this tension.

Guiding Principles

After several discussions amongst community members about scoping work in FY2011, a number of key guiding principles are emerging:

- The best focus for the CDI leadership team is on facilitating relationships within the community and interconnection between existing projects. Trying to fund major development activities is not the best use of any available CDI funding.
- The community needs to facilitate much more direct interaction between data practitioners and research scientists to make sure scientists know about available tools and methods and that technologists are focused on the highest priority problems. Some of this can be facilitated through interactions at the strategic level with SSPT teams, but much more can be done in small and large settings with on the ground science projects.
- Conference calls and WebEx sessions are no substitute for in person, focused workshops where much more collaboration and discovery of shared interest/talent can be done. The cost in time and money of face-to-face meetings will be more than compensated for in faster overall "time to market" for new capabilities.
- The community started some good work in 2010 but needs to take those products all the way to production release so that they can be used by all USGS scientists. This includes both technology development/release as well as a training and consultation program to provide direct data management assistance.

Community of Practice Facilitation

Earlier iterations of how to put the 2011 proposal together focused much more heavily on the production of specific tools and components for data integration. Discussions amongst various teams let the group much more in the direction of targeted facilitation of the Community of Practice Working Groups. This was described by one community member (paraphrased here) as giving teams with generally parallel goals and objectives a reason to get together, discover the interdependencies between their projects, and coming up with ways to make a stronger and more broadly applicable end product.

Examples of this include:

- Getting the GDP Tools group with the Data Uploader group to work out how the Uploader product suite will handle, serve, and document data sources (mostly NetCDF and WCS) of for use by "data integration portals" built on the GDP Tools.
- Getting the Metadata Working Group together with the Technology Stack Working Group (TSWG) to work on metadata standards and profiles to be served up via the CSW service on the technology stack.
- Making sure that community members heavily engaged in working on Data Management Best Practices and the Metadata Working Group can be involved in Coastal and Marine Spatial Planning activity in 2011 so that the community as a whole benefits from that important focus area and the CMSP activity gets the crosscutting talent it needs to complete a robust plan.
- Making sure that the Science Strategy Planning Teams have input from the Community of Practice to make sure that their plans include the necessary elements to identify data integration goals and, conversely, that science questions from those teams make their way into the Community to inform priorities and actions.
- Focused training, consultation, orientation, and fact-finding sessions between data practitioners/technologists and research scientists to provide data management support and discover "ground-level" needs.

These concepts need to be more thoroughly fleshed into some set of discrete actions and funding amounts for the forthcoming proposal.

Documentation and Outreach

Peter Fox (RPI), who addressed the Community at the August workshop, made a statement about appreciating the open attitude of our USGS community but indicating the fact that he, as an outsider, was not able to collaborate in our online space. We took this to heart and would like to propose the following:

- Creation of a curated Community for Data Integration Web space with a form and structure published as an official living product (may require pushing the limits on the USGS publications process) that can serve as an official documentation point for actions and outcomes from the Community. This will require a working group to be spun up at least temporarily to lay out this structure, assemble initial content, and work out a viable long-term publishing process.
- Migration of the current myUSGS wiki and document space to a more easily accessible forum where external users with interest can be granted access in a simpler manner. Some attention should be paid to the form and structure of this open space as well, albeit with less rigor than the curated space, and a relationship may exist between the two spaces.

Publishing Products

The 2010 work produced several good products that are candidates for publishing into the open source marketplace. Publishing as such will a) provide an official outlet for the software and documentation and b) open the products up for collaboration by the broader community. Several community members participated in an exercise to rewrite a USGS policy that will support this type of software release, but there are still a number of challenges in putting open source publishing into practice (e.g., publishing processes, Center and Bureau approval, Fundamental Science Practices, etc.). It is proposed that some level of funding and energy be put toward releasing one or more products in 2011 through an appropriate open source venue, working through all of the necessary practices in the USGS. These practices would be codified and released through the CDI Web space (discussed above).

> The "toolkit" items below were one attempt to discuss FY2011 work that focused more on producing tools and widgets than on facilitating community interactions. These concepts and associated goals are still valid and will be part of the overall work, but the focus for the proposal has shifted toward community of practice facilitation and accomplishment of these goals within that framework.

Data Integration Toolkit (DI-Toolkit)

This element will essentially begin building on FY2010 work on the Geo-Data Portal (GDP), the ArcGIS tools for data access, and other aspects of directly applying Web services access to data to produce an overall toolkit that can be plugged into various science questions in need of integrated data.

Possible Leadership Team Identified at the Workshop: Carma San Juan, Roland Viger, Bruce Jones, Mike McHale, Robin O'Malley

Major Elements

- Identification, documentation, and broad availability of Web-based, open standards tools for data integration
- Identification, documentation, and broad availability of the ESRI-specific tools for data integration
- Application of the DI-Toolkit to one or more major science questions through partnership and responsiveness to the Science Strategy Planning Team (SSPT) process or another major project identified through the CDI. This should include a final report published as a USGS series or submitted to an outside journal that documents important technological aspects of the DI-Toolkit and demonstrates how the tools were employed.

Continuation of FY2010 Work

- Much of the basis for this concept came from work by the FY2010 Goal 1 and 2 teams on the Geo-Data Portal. John Hollister and Dave Blodgett came up with the idea of calling that the "GDP Toolkit" based on the direction they went with the architecture as a baseline enabler of many different portals, using all of the geoprocessing and data integration capability built into the platform. Work under the DI-Toolkit concept in 2011 will help to put the GDP Toolkit into full production.
- This overarching goal encompasses the ongoing work started late in the year by Brian Reece and Sally Holl (TX-WSC) to develop a Web Services-based ArcGIS toolkit for accessing NWIS and retrieving/building a local GeoDatabase. This is envisioned as an important part of the DI-Toolkit targeted toward ESRI users.
- Peter Schweitzer also began working with the Service-based tools developed by David McCulloch in relation to the Web services provided by the Mineral Resources Data System. This may develop into a tool in the kit for MRData.

Original Workshop Bullets (working points)

- Science Questions for Application of the Toolkit
- ~~Move to Tech Stack Serve TNM, GloVIS, MRDATA, NWIS (streamflow and water quality), watershed boundaries at National scope with Statistical Capability~~
- Compile scientific question(s) for testing each service (e.g. qualifying affect of land-use change and climate change on streamflow change)
- Geoprocessing Services – Curtis Price
- Use of the Repository Virtual Appliance for hosting select technologies in the toolkit - Tim Kern

Data Services Toolkit (DS-Toolkit) a.k.a. Tech Stack Working Group

- Dedicated wiki page for the Technology Stack Working Group

The wiki page listed above should be consulted for the most recent developments of this group. The following text reflects thoughts at the time of the workshop, which continue to evolve. In general, this element builds on the FY2010 work for the Data Uploader and the concepts discussed heavily in the workshop about a group of data hosting and serving technologies operating in different parts of the USGS. The hope is that all these developments can be coordinated to form a "Scientific Data Network." At the very least, this element seeks to develop a community of practice around the technology and design of such a network.

Scientists may opt to use a centralized incarnation of the Data Uploader or download and deploy a local version of this platform (this version is sometimes referred to as a "droppable appliance") within their project. In addition, the design of this platform and the community discussions around it will help to provide a template and guidance for projects that need to develop their own information management system ("appliance") and will hopefully result in a project-specialized system that can still participate in the greater Scientific Data Network.

An emphasis that has grown since the workshop that is worth mentioning here is that this community seeks to develop understanding and recommendations on how such appliances can be engineered to effectively exploit the many corporate/enterprise data services, such as those being pushed forward by the National Map and the National Water Information System. While the Data Uploader has consciously attempted to build around open-source software (in addition to open standards), many of these corporate services do leverage important proprietary components of the Agency's corporate computer model, such as ESRI ArcGIS Server.

Possible Leadership Team Identified at the Workshop - Blodgett, Viger, Dadisman, Kern, Gunther, Skinner, Hope/Tricomi, Greenlee

Major Elements

- Develop the Tech Stack Working Group.
- Engage the developers of major corporate components, including the Data Uploader, TNM services, and NWIS services with this group.
- DS-Toolkit engineered, documented, and established for nodes targeted on current hosting needs from the Science Strategy mission areas and specific priorities. These will include, at a minimum, a node for the Coastal and Marine Geology Program (focused toward Coastal and Marine Spatial Planning), nodes for the Climate Effects Network (Upper Colorado River Basin, Greater Platte River Basin, Yukon River Basin), and node (s) for the National Climate Change and Wildlife Science Center.
- Coordinated Catalog Services (probably based on the Open Geospatial Consortium Catalog Service for Web (OGC-CSW) standard) and a central harvest of all nodes in the DS-Toolkit network to support broad search and discovery.
- Application of the DS-Toolkit to one or more projects/data hosting needs identified for the nodes and targeted to a major science question. Data services from the DS-Toolkit should feed directly into usage from the DI-Toolkit. This application of the technology stack will be released in a USGS series report or submitted to an outside journal to both document the approach and demonstrate its applicability.

Continuation of FY2010 Work

- Work will build on the approach piloted by the Data Uploader group to create a "droppable virtual machine (VM)" with a suite of supported technologies that will advertise/register hosted data resources with a shared metadata cataloging service.
- The workshop teams responsible for this element discussed an approach that would embrace both open-source methods and technologies along with the elements of the ESRI technology stack that serve open standards. This builds on discoveries made about the relative efficacy of these two approaches.

Original Workshop Bullets

1. Develop a Technology Stack Working Group
    a. Unified Access Framework for Gridded and Vector Data: WMS, WCS, WFS, and OPeNDAP+CF - Rich Signell, Blodgett, Viger
    b. Discoverability and Interoperability of Web-based Data and Processing Services - Blodgett, Viger, Dadisman, Kern, Gunther, Skinner,
    c. Catalog Services Plan, including architecture, as well as search and harvesting functions. With metadata standards - Steve Richard## Web Application Integration Framework (Summary Notes Presented) - Matt Tricomi, Greg Smoczyk, Hollister
    d. ~~Moved from DI Toolkit~~ to Tech Stack Create USGS Corporate Services Integrated Roadmap for serving TNM, GloVIS, MRDATA, NWIS (streamflow and water quality), watershed boundaries at National scope with Statistical Capability (TNM, WBD, NWIS, MRDATA, GloVIS)
    e. Data/Application Publishing - Web Services Publishing Best Practices (TNM, WSWG) (Moving from DM to Tech Stack since about publishing tech services and to allow NGP to participate in one group - Discussed at 12/7 TS meeting)

Data Management Toolkit (DM-Toolkit)

The success of the fully applying the DI-Toolkit for scientific applications and using the DS-Toolkit for hosting and serving usable data will depend heavily on the organization's ability to effectively manage its data and associated documentation (metadata). (Reference the Greenlee User Story.) Building on the "PI-Toolkit" concepts discussed in relation to documenting data uploaded through the Uploader during the 2010 work, a much larger team met to discuss the overall dynamics of managing, documenting, and publishing data. This group examined social and organization dynamics inherent in the data management process and identified a number of areas for forward momentum in 2011.

Possible Leadership Team Identified at the Workshop - Heather Henkel, Viv H., Adrian, Huffine, Dadisman, Frame/Mancuso, Kase/Fornwall, Schweitzer, Govoni

Major Elements

- Bureau Metadata Technical Support Team - Support System/Infrastructure: documentation, helpdesk, in-house consulting services
- Metadata Creation and Management Framework
- Application of the metadata technical support team, applicable data publishing policies, and data management practices to data hosted and served through the DS-Toolkit and used through the DI-Toolkit to address one or more major science questions. This work will be released in a USGS series report or submitted to an outside journal to document the relationship between policies and practices and the use of technical support staff in the effective handling and release of major data products.

Original Workshop Bullets

- Data Management Best Practices… to protect data and applications for posterity
    1. Aggregation of Existing Metadata to Support Science Programs
    2. Future of Enterprise Geospatial Technical and Scientific Support Under Data Integration Theme
- FY 2010 wrap-up: finish, market, and transition into production
    1. Develop end-to-end toolkit/workflow for uploading, nurturing metadata, curating, etc. - Roland Viger and Tim Kern

Existing Data Management Tools and Products as Practical Examples

- Burley, T.E., and Peine, J.D., 2009, NBII-SAIN Data Management Toolkit, U.S. Geological Survey Open-File Report 2009--1170, 96 p. Available at http://pubs.usgs.gov/of/2009/1170/ (teburley@usgs.gov, 20101028)

Knowledge Management Framework (KMF) and Toolkit (KM-Toolkit)

Establishment of a formal, well-structured, and comprehensive Knowledge Management Framework (KMF) to organize, presrve, and share information and resources relevant to USGS scientific data integration and management – both resulting from, or referenced in the course of CDI projects – should be an essential component of the CDI effort. Such a framework would provide a ready and essential base for the documentation of systems, tools, standards, processes, and practices which could be easily drawn upon in support of CDI-related community learning, training, technical support, and general communication efforts.

We propose that a workgroup be formed to address the need for creating and managing a sound KM Framework for the benefit of the CDI and its Science partners.

A Knowledge Management Framework page has been created to flesh out the initial ideas summarized here:

Information resources falling under this framework might include:

- A version-controlled "digital library" of documentation, educational, and general communications materials
- Various registries or catalogs providing pertinent information about available tools, services, or discovery resources to encourage use and re-use
- Other relevant "knowledgebases" (KBs)
- General shared knowledge resources for collaborative learning

Tools to support KM would include:

- Internal/extranet "working" wiki(s)
- Public public-facing "curated" wikis – for general knowledge sharing
- Shared citation and bookmarking services
- Formally "information architected" internal, public, or quasi-public website
- Concept, Mindmap, or similar brainstorming and modeling tools

Workgroup membership: Dave Govoni (convenor), Richard Huffine (co-convenor)

Funding Considerations and Logistics

- One of the primary considerations discussed at length during the workshop was the need to ensure proven applicability of technologies and methods developed through the CDI to core "big science questions" posed through the Science Strategy planning process. To this end, the CDI is working to:
  - Identify key scientists to help articulate questions and work with technologists to document the connections to the Toolkits discussed in 2011 planning.
  - Fully articulate several major science questions to be addressed by the Toolkit approach and successfully apply the tools to at least one of these during 2011.** Participate heavily in the Science Strategy Planning Team (SSPT) process, not only for Core Science Systems but the other mission areas as well. This will help a) provide focus for data integration activities, b) feedback useful strategic considerations to all SSPT plans, and c) establish a long-term approach for the direct connection of science and technology to achieving the USGS Science Strategy goals.

- The CDI projects could greatly benefit from the focus and discipline of producing both final report products and publicly released open source software. Both of these approaches introduce useful scrutiny and independent peer review of the work, ultimately resulting in more robust products. These processes do require additional time and funding, both considerations for the focus and scope of the 2011 proposal.
- One of the lessons learned in 2010 was that effective coordination across such a diverse group of scientists, technologists, and data practitioners requires more face-to-face communication. Regular weekly check-ins were good, but the 2011 teams need 3-4 focused, in-person meetings throughout the year. Project funds will need to be made available to facilitate these workshops.
- Both the DI-Toolkit and the DS-Toolkit include software components requiring servers and other hardware. Across the Community for Data Integration, the USGS has a tremendous wealth of technological knowledge and expertise to engineer, operate, and maintain these systems. The overall approaches discussed of releasing adapted and developed software as open source projects and engineering VMs to contain a particular technology stack are both conducive to distribution of well-managed elements to systems administrators close to projects and teams where the toolkits will be applied to science questions. Combined with this, however, the Department of the Interior and Office of Management and Budget are both pushing heavily toward physical installation consolidation under both environmental and cost considerations. To this end, the teams will be looking toward ways to provide a cloud computing platform within the USGS network where the best physical locations can host multiple VMs managed by Science Centers throughout the USGS.

Research Resources

- Ray Obuch will post 1998 BLM Data Management Workshop results as a model for ...

Comments

DBlodgett's high level observations from 9-2-2010 call:

- Publication needs to be suitable to the content. High level-abstract stuff might go in a formal publication, operational documentation and instructional lessons might go in a wiki. We shouldn't force anything that doesn't fit. We need to reward the creation of good "documentation" (for lack of a better word) whatever form that takes.
- The "science questions" are not huge overarching things, but rather they are small project level issues that represent larger issues.
- We need to collaborate across disciplines/projects with face to face meetings (sponsored by CDI)
- Some issues are not going to be solvable via technology. (ie. the publication process doesn't do data management...really)
- The data management/best practices/metadata practitioners need to be represented at face to face meetings as guidance and sounding boards.

- We need a diverse set of people in the room to act as advisors and sounding boards at workshops. This diverse set of people can offer accurate guidance in their areas of expertise. In order to know how to enable/integrate our data for other communities those communities need to be at the table.

Hi,

Couple of comments/questions:

- I appreciate the need to title the different parts, but I find the shorthand of "DI-Toolkit", "DS-Toolkit," and so on a bit confusing. I'm worried that this might impact the readability of our proposal. I don't have any alternates to suggest, unfortunately.
- With regard to the KM group:
    - I think it might be important to clearly distinguish this new topic from pre-existing ones. There seems to be a ton of overlap with the Data Management group, in particular. While overlap is not inherently bad, perhaps Dave/Richard could talk w/folks from this other group to make this separate clearer to reduce the impression of redundancy.
    - Also wondering about how project or domain-specific portals relate to the KMF. I imagine that more focussed efforts will get scientists' attention first and that their products might then be migrated or federated with a larger KMF infrastructure. Do I have this right?
    - Is the KM group talking about development and funding for 2011 or is this a longer-term concern that we will address once a few more pieces of the various toolkits materialize and start getting used?

Thanks!

--rviger@usgs.gov, 14-Sep-2010