Statement of Interest

Title:
Furthering the Implementation and Practice of Data Management in the USGS Chesapeake Bay Studies

Description:
Expand the Eastern Geographic Science Center's Research and Development Computing Cluster (RDCC) with applications to support data integration and local scale data management among USGS Chesapeake Bay studies science members. Additionally, dedicate more support to projects to educate on data management best practices and to provide a data steward.

Region:
Eastern Geographic Science Center

Project Leads:
Cassandra Ladino
12201 Sunrise valley Drive
MS521
Reston VA, 20192
Phone: 703-648-6188
Fax: 703-648-4603
Email: ccladino@usgs.gov

William Miller
12201 Sunrise valley Drive
MS521
Reston VA, 20192
Email: bmiller@usgs.gov

Partnerships and Communication:
This project brings together the multiple science centers that make up the USGS Chesapeake Bay studies and the computing technology at the Eastern Geographic Science Center with additional data management work flow expertise from CDI members in the National Climate Change and Wildlife Science Center (NCCWSC) and Climate Science Centers. The USGS Chesapeake Bay science coordinator, Scott Phillips has identified fish health and contaminants as a priority for this year's data management activities. Cassandra Ladino (EGSC) will act as a data facilitator and work closely with Vicki Blazer (Leetown, WV SC) and Patrick Phillips (NY WSC) to load, integrate, and provide applications for their data on EGSC's Research and Development Computing Cluster (RDCC) developed and maintained by William Miller. Cassandra Ladino will also continue and improve her current role as the Chesapeake Bay studies data steward by consulting with Emily Fort (NCCWSC) and drawing on the real world expertise of the NCCWSC in data curation and development of data management plans. The NCCWSC was specifically selected based on their similarity to the broad science topics covered by the USGS Chesapeake Bay studies and their maturity in successfully implementing

data management into project workflows. This project brings together subject matter of interest to both the CDI Tech Stack and Data Management working groups and will provide content valuable to the members of each community.

Project Summary:
Collective USGS efforts in the Chesapeake Bay watershed began in the 1980s, and by the mid-1990s the USGS adopted the watershed as one of its national place-based study areas. Great focus and effort by the USGS have been directed toward Chesapeake Bay studies for almost three decades. The USGS plays a key role in providing science to improve the efficiency and accountability of Chesapeake Bay Program activities. Each year USGS Chesapeake Bay studies produce published research, monitoring data, and models addressing aspects of bay restoration such as, but not limited to, fish health, water quality, land-cover change, and habitat loss. The USGS is also responsible for collaborating and sharing this information with other Federal agencies and partners as described under the Presidential Executive Order 13508 Strategy for Protecting and Restoring the Chesapeake Bay Watershed signed by President Obama in 2009. The USGS Chesapeake Bay studies recognizes the importance of all encompassing project level data management and began coordinating in 2011 with the EGSC and the CDI to investigate implementing comprehensive data management solutions to facilitate data sharing and the development of decision support tools.

Facing many challenges, the implementation of data management into the workflow of USGS Chesapeake Bay studies projects has been little more than a pilot study with additional theoretical planning. The Chesapeake Bay studies' greatest challenges are finding dedicated and specialized resources to provide to projects to support education, data management planning, and data curations. These tasks are too much of an initial burden on projects and their utility must be demonstrated before adoption will happen at the project level. This collaboration between the Chesapeake Bay studies' fish health and contaminants project and EGSC's RDCC, along with expertise from the NCCWSC, aims at providing a full data management workflow implementation example that will demonstrate capability and utility.

At a project level, scientists are most interested in data integration and data analytics. While data storage and sharing are also important, they are typically not prioritized among the most important aspects of a scientific investigation. The Chesapeake Bay studies have used available USGS-wide computing infrastructures and data catalog applications focused on storage and sharing to meet project data management needs to the greatest extent possible, however, existing tools are not well suited to satisfy the goals of scientists. As a result, we propose to expand EGSC's RDCC to provide an infrastructure where scientist can integrate multidisciplinary data sets and run data analytics. The RDCC is a persistent and highly scalable system that is currently equipped with OGC map, coverage, and feature services, relational database software (Postgres), a data repository service (Fedora and ESCIDOC), and a Hadoop Cluster. The RDCC's current functionality is to store full text research reports, satellite and other imagery, as well as scientific data. While the RDCC is not fully tailored to Chesapeake Bay study needs, it does provide a robust and open architecture upon which we can build databases to integrate multidisciplinary data and develop the necessary applications to load, access, and analyze data.

The proposed project can be divided into three sequential phases: planning and education, database assembly, and application development. The planning and education phase primarily focuses on coordination with the NCCWSC to apply their data curation and data management plan development techniques to the USGS Chesapeake Bay fish health and contaminants projects. We will also provide this use-case information to other projects within the Chesapeake Bay studies through presentations to begin educating the entire group. The results of this phase will be a data management plan that can be referenced for future year work and an understanding of the fish health and contaminants project data holdings. The database assembly phase primarily involves developing a Postgres database and schema to store fish health and contaminant data on the RDCC. This phase builds off the understanding of the fish health and contaminants data gained in the previous phase. Additionally, however, we will define relations among the data with the expertise of the project scientists to ensure accurate schema development. Once the database schema is defined, the data is loaded, and a series of initial manual queries are made to test the database, the application development phase will begin. The application development phase will be focused on providing better access to the RDCC server and easier interaction with the database, so that the data base administrators and data stewards do not have to handle every data request desired by the scientists. Development will be light in the first year, but provide a foundation to build upon. In this phase also, we will work closely with the scientists to develop use cases. While this is a substantial amount work, we have many of the components to make this project successful already available to us and can take advantage of them with increased support. Ultimately, demonstrating the utility of data management in this use case will provide significant advancement in adopting data management into the workflow of all Chesapeake Bay studies projects.

Cost:

Items	Description	In-kind	Requested
RDCC Server Support	Maintain and install components		3,000
Database Development	2 database administrators (1/4 year)	25,000	25,000
Application Development	1 Senior Developer (1/4 year)	15,000	10,000
Data Steward	1 Physical Scientist (1/2 year)	20,000	10,000
Project Management	1 Supervisor (less than 1/4 year)	10,000	
TOTAL		**70,000**	**$48,000**