USGS
science for a changing world

Defining a Data Management Strategy for USGS Chesapeake Bay Studies

Open-File Report 2013–1005

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

Defining a Data Management Strategy for USGS Chesapeake Bay Studies

By Cassandra C. Ladino

Open-File Report 2013–1005

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

U.S. Department of the Interior
KEN SALAZAR, Secretary

U.S. Geological Survey
Marcia K. McNutt, Director

U.S. Geological Survey, Reston, Virginia: 2013

Suggested citation:
Ladino, C.C., 2013, Defining a data management strategy for USGS Chesapeake Bay studies: U.S. Geological Survey Open-File Report 2012-1005, 7 p. (also available at *http://pubs.usgs.gov/of/2013/1005/*).

Contents

Figures

Defining a Data Management Strategy for USGS Chesapeake Bay Studies

By Cassandra Ladino

Introduction

The mission of U.S. Geological Survey's (USGS) Chesapeake Bay studies is to provide integrated science for improved understanding and management of the Chesapeake Bay ecosystem. 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 using "ecosystem-based adaptive management, which will provide science to improve the efficiency and accountability of Chesapeake Bay Program activities" (Phillips, 2011). 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 responsible for collaborating and sharing this information with other Federal agencies and partners as described under the President's Executive Order 13508—Strategy for Protecting and Restoring the Chesapeake Bay Watershed signed by President Obama in 2009. Historically, the USGS Chesapeake Bay studies have relied on national USGS databases to store only major nationally available sources of data such as streamflow and water-quality data collected through local monitoring programs and projects, leaving a multitude of other important project data out of the data management process. This practice has led to inefficient methods of finding Chesapeake Bay studies data and underutilization of data resources. Data management by definition is "the business functions that develop and execute plans, policies, practices and projects that acquire, control, protect, deliver and enhance the value of data and information." (Mosley, 2008a). In other words, data management is a way to preserve, integrate, and share data to address the needs of the Chesapeake Bay studies to better manage data resources, work more efficiently with partners, and facilitate holistic watershed science. It is now the goal of the USGS Chesapeake Bay studies to implement an enhanced and all-encompassing approach to data management. This report discusses preliminary efforts to implement a physical data management system for program data that is not replicated nationally through other USGS databases.

Data Management System Services

The foundation for any data management activity is typically a data management life cycle, a diagram that shows the transitional stages of data as they move from raw or new data to preserved and published data. The USGS Chesapeake Bay studies have been working closely with the USGS Community for Data Integration (CDI) to incorporate their scientific data life cycle model that has become the accepted USGS model into the Chesapeake Bay studies data management work. CDI's scientific data life cycle model has six stages: Plan, Acquire, Process, Analyze, Preserve, and Publish/Share, with three main cross-cutting themes, documentation, quality assurance, and data security, applied at each stage (Data Management Working Group, 2012) (fig. 1).

The scientific data life cycle model does not represent the same concepts as the steps in the traditional scientific method, but the data life cycle can be thought of as a parallel subcomponent of the scientific method. The scientific data life cycle model focuses on the evolution of data as they move through the steps in the scientific method and does not focus on the process of conducting science in general. The terms used in the scientific data life cycle model do not refer to terms used in the scientific method but rather to core stages in data collection, scientific data investigation, and data distribution. Key terms associated with the data management system are described in the following sections.

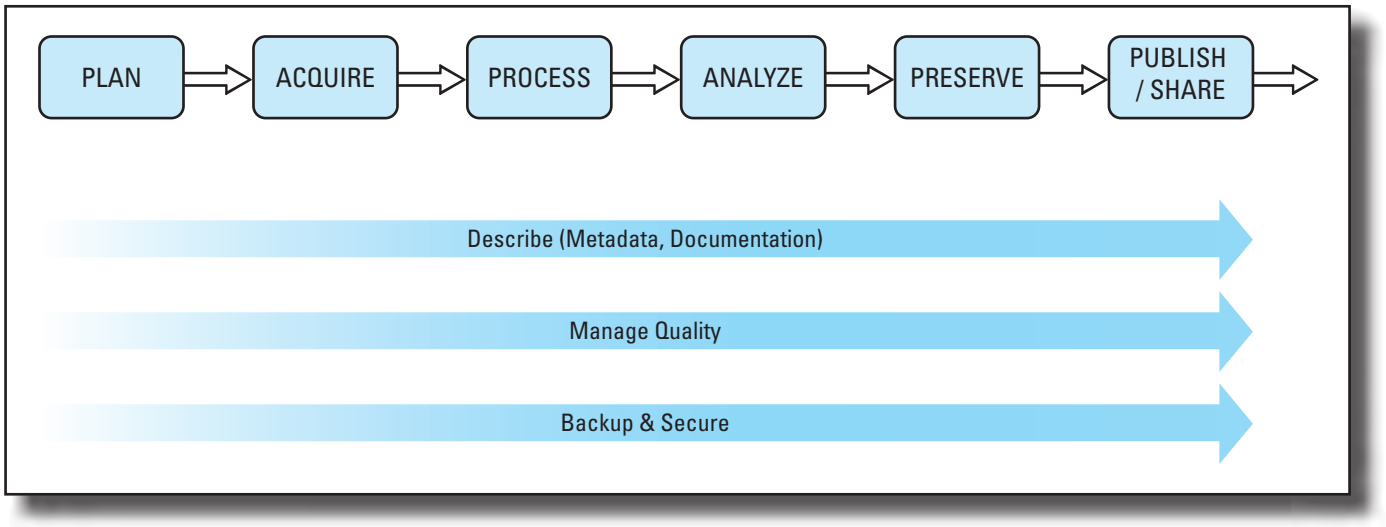

Figure 1. Community for Data Integration scientific data life cycle model (Data Management Working Group, 2012).

The data management system for USGS Chesapeake Bay studies has to accommodate many stages of the scientific data life cycle model and a variety of audiences such as USGS data producers (scientists and technicians), external partners (State, Federal, local, and nongovernmental organizations), and internal or external Web applications. As a result, three main user services were considered: data uploading, data searching, and data as Web services for applications. Preliminary efforts take advantage of an existing organization-wide resource, the USGS ScienceBase, an online data cataloging application and platform for collaborative data management (Fort Collins Science Center Web Applications Team, 2012b). Besides being a structured database containing metadata and physical data, ScienceBase provides specific tools for data upload, registry, and access (Fort Collins Science Center Web Applications Team, 2012a) that enabled the Chesapeake Bay studies to meet project goals. ScienceBase was chosen for this project because it is Web accessible, provides a robust set of services, and allows broad USGS-wide connections to be made with USGS Chesapeake Bay data (Fort Collins Science Center Web Applications Team, 2012b).

Data Uploading

Data uploading, a graphical user interface for collecting well-documented scientific data and analysis results, needs to be designed to support the Preserve and Publish/Share stages of the scientific data life cycle model and primarily the USGS research scientist community. The CDI model describes Preserve as the "actions and procedures to keep data for some indefinite period of time," and Publish/ Share is described as the process "to prepare and issue, or disseminate, the final data products of the research or program activity" (Data Management Working Group, 2012). To meet these requirements, a data uploading tool needs a database structure and database management system software to store, organize, and provide access to datasets. Additionally, an intuitive nontechnical graphical interface is needed to allow interaction between data producers and the database. This is a key component, as the interface should not overburden the scientist or technician and should be designed to integrate intuitively into a project workflow. Making the interface Web accessible is a good approach to ensuring ubiquitous access to all users; however, the end result may be less integrated into individual workflows. Lastly, the interface needs to collect metadata along with the science data. There are two common approaches to collecting metadata; some interfaces allow the user to upload a standard metadata file that follows a specified standard such as Federal Geographic Data Committee or International Organization for Standardization. Another approach is to allow users to fill out text fields that ask for pertinent information such as author, description, and date in an informal manner. The formal approach has the advantage of providing great detail and may already be a part of the research scientist's workflow, while the informal approach can be faster for researchers who have data that do not need formal metadata. Ultimately, the convenience of the data producers should be the deciding factor if the data do not require formal metadata.

The ScienceBase Web interface contains a file uploader and data documentation wizard (fig. 2). ScienceBase uses the informal approach of having the user fill out key information pieces rather than upload formal metadata files. The user is prompted to step through each tab to fill out the metadata information and eventually upload the dataset. The result is a searchable record for the data (fig. 3).

While ScienceBase is Web accessible to all USGS employees, for the Chesapeake Bay studies, one data steward has been identified to be responsible for uploading the most widely used datasets in the initial years. This approach allows progress to be made in adding data to the data repository and to create a better informed data management plan that will be tailored to the workflow of USGS data producers. It is a project goal to have a lead scientist from each project to ultimately be responsible for uploading key datasets that are not replicated nationally through other USGS database systems.

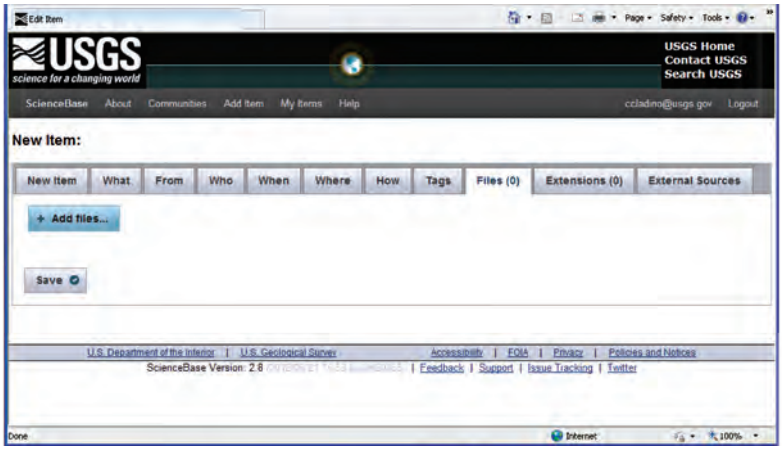

Figure 2. ScienceBase Uploader graphical interface.

Data Searching

A data search engine, a graphical interface for finding and obtaining copies of datasets in the data repository, needs to support the Acquire and Publish/Share stages of the scientific data life cycle model and two important communities, USGS scientists and the external Chesapeake Bay Program partnership community. Acquire is described by the CDI life cycle model as the process that "includes automated collection (for example, of sensor-derived data), the manual recording of empirical observations, and obtaining existing data from other sources" (Data Management Working Group, 2012). The components of a data search engine needs to include functionality for querying the data repository and functionality for allowing users to obtain a copy of the stored data. Supporting the Acquire stage is the primary objective of the data search engine, as it makes data available for download. To a lesser extent, a data search engine also supports the Publish/Share stage as described above, because the audience must be made aware that the data exist before the data can be shared and downloaded. It is difficult to design a user interface for a data search engine, because the audience tends to vary greatly in knowledge, needs, and habits. In this case, it is most effective in the initial project stages to only consider USGS scientists and Federal, State, local, and nongovernmental partners doing science in the watershed as users because they are a large primary target group with similar expertise, yet still provide the opportunity to develop an interface for a user group with diverse needs and habits. The search interface should be minimalistic in design but include the most common ways to search for data such as data type, text, and location. The functionality and vocabulary of the interface should follow community standards rather than be USGS specific.

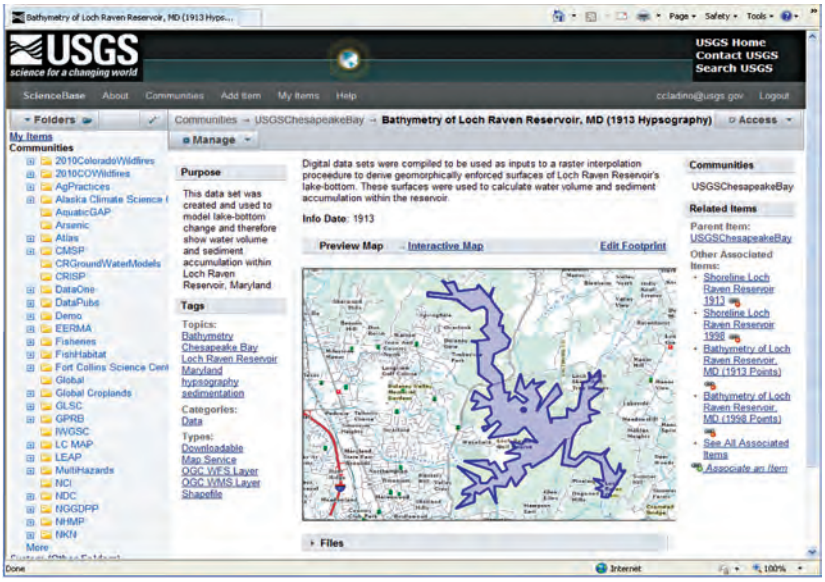

Figure 3. Example data record in ScienceBase.

The ScienceBase catalog allows search through the main portal, which includes all community-contributed datasets, or search through specific community portals. The USGS Chesapeake Bay community search portal provides search by data category, text, and location (fig. 4).

The USGS Chesapeake Bay ScienceBase community currently contains 19 downloadable datasets that represent the most comprehensive and popular products, 31 relevant Web site links, and 314 Chesapeake Bay related publications.

Lastly, ScienceBase has an Application Programming Interface that exposes record-level information about datasets through Web services. These Web services have been used to create a custom data search interface for the USGS Chesapeake Bay Studies Web site (*http://chesapeake.usgs.gov/*). The custom portal is a handy tool to allow partners to browse the scope of the datasets in a simplified manner (fig. 5) and in the future data will be able to be downloaded directly from the portal. The custom search interface is location driven, showing data points on the maps and then listing the datasets at each point when the user clicks on the map.

Data as Web Services

Data as a Web service comes out of the need for customers to have access to a dataset but not to own a physical copy. Data as a Web service has many advantages for customers, such as eliminating the need for dedicated hardware space, reducing redundancy, and maintaining data quality. The premise is that the owner of the data maintains the data and provides a Web-accessible and persistent link to the data that the customers can consume through a variety of methods to enable them to view the information contained in the dataset. Common Web services include Web Map Services, Web Feature Services, and Simple Object Access Protocol. These types of services are often coupled with other services such as Open Geospatial Consortium Web Processing Services to allow the data customer to do basic "on the fly" analysis and manipulation. Data as a Web service is an alternative source for project data and an alternative method for publishing and sharing project data, and therefore should be considered at all stages of the scientific data life cycle model. Web service customers range from highly technical Web developers and programmers to Web applications. While Web service customers may seem like a small niche in the larger customer community, this is an important group because these customers promote re-use and increase exposure of the data products. When designing Web services for data in the data repository, it is best to follow open standards, which are interoperable with the largest number of applications, and to provide as many different types of Web services as is feasible.

ScienceBase allows a Web developer or Web application to request and view a data record's footprint and attribute information in a variety of Web service formats. ScienceBase also provides Web service end points to the actual spatial datasets. This function allows for an effective and streamlined approach of having the data management system feed other Web applications to ensure that they are serving consistent and up to date information. The results of this function are to great advantage as the USGS Chesapeake Bay studies develop future Web applications for decision support.

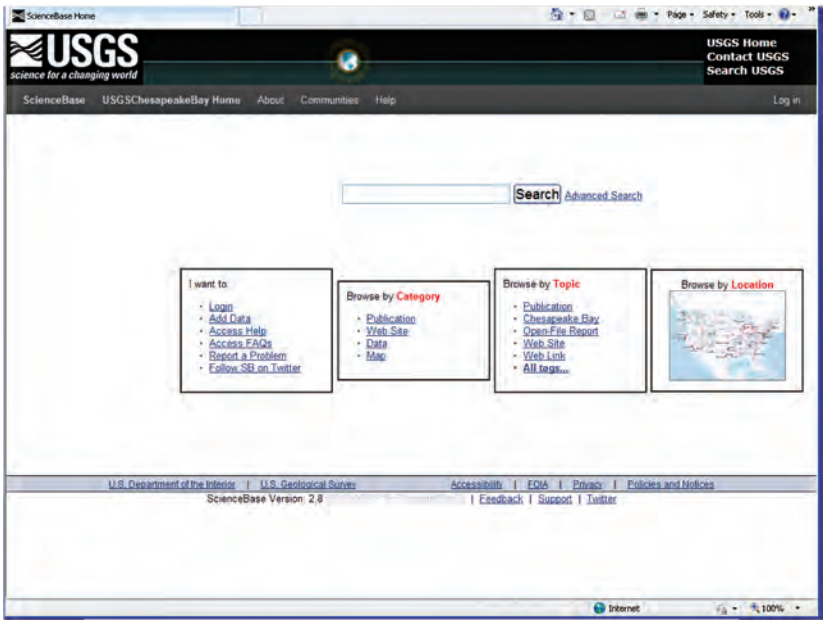

Figure 4. ScienceBase graphical interface for searching for data.

Figure 5. Custom USGS Chesapeake Bay activities portal to ScienceBase.

Implementation Challenges

Even though the data management strategy is in the early phases of development, many challenges are evident. Some of the largest issues that need careful planning involve changing roles and responsibilities, minimizing redundancy, and ensuring the future stability of the data management system. While these challenges are common to many groups implementing a data management system, below are some examples of how they specifically apply to the USGS Chesapeake Bay studies.

Changing Workflows

All employees, including student and contract hires, will have to adhere more closely to the requirements of each stage in the scientific data life cycle model, adding new workflow tasks and changing how and where they store data. Data will have to be handled more formally from acquisition to publication and will require documentation at each stage in between. In a simplified version of the current workflow process, it is common for one employee to collect the data, for another or maybe the same employee to conduct some analysis, and then for subsets of the collected information to be summarized in a publication with little documentation as to the transformation of the data in between collection and publication. A new workflow process tailored to supporting

data management would require the collected data to follow a standard convention for naming, units, and structure. The process would also require some minimal information about how and when the data were collected to be documented and distributed with the data. The analyst would then be responsible for adding a summary of the methods to the metadata, such as filing and packaging any new data files with the original collected data. In this manner the data can be a stand-alone product along with the typically resultant publication.

To successfully integrate data management into workflows, traditional concepts of roles associated with employee job titles and measures of success will need to be challenged. For example, the scientist's measure of success is traditionally coupled with writing publications. Research activities are structured to facilitate the production of publications. To encourage integration of data management as part of the research function, the measures of success for scientists will need to reward for not only publishing research findings but also for making data accessible to customers – in other words, publishing stand-alone datasets. Although there are many possible incentives such as more reward opportunities for the scientists and immensely increased exposure of USGS science, overcoming traditional measures of success will be difficult. Additionally, new job titles may need to be created and roles may need to shift. Data producers may find themselves working more closely with and sharing responsibilities with others such as Web developers, who have not historically been a part of their workflow. Changing employee roles and measures of success are essential for the successful integration of data management into workflows, but the process will be slow because of the great number of personnel involved. The USGS CDI has already begun to discuss some of these issues. The results of those discussions should ultimately help facilitate change at the program levels.

Integrating External Data Stores

Much of the science that is performed in the Chesapeake Bay studies utilizes partner data as well as data collected and produced by the USGS; therefore, it is necessary to account for partner information. Partner data can come from a variety of different sources such as other databases, publications, and Web sites. It is important to account for these data, as they are an integral part of USGS science. Some would argue that having copies of partner data is the only way to ensure real preservation because databases and Web links tend not to persist over time. The USGS cannot possibly maintain copies of all partner datasets, nor is it desirable, for two reasons. First, maintaining these data would quickly overwhelm USGS resources, and secondly, this scenario would create redundancy and possible data inaccuracies if there is a lack of a well-defined plan for data transfer. Even within the USGS, there is a need to consider how to incorporate subsets of data from larger systems such as *The National Map* and avoid duplication. In the current ScienceBase implementation of a data management system, the decision was made to create records for URLs that point to partner Web sites with data or Web services providing data. This is a minimum effort approach, but it is not entirely comprehensive or seamless. Much effort would need to be put into enabling partner data stores to be harvestable and to have the ability to transfer their collection of metadata records to another data catalog such as ScienceBase. ScienceBase has the capability to harvest other data catalogs, such as the USGS Publications Warehouse, but currently, use of this option has not been explored with any partners.

Long-term Persistence

It is easy to take for granted the longevity of a server or Web application, but there are many factors that may shorten the lifespan of an application such as a data management system. Data continuity involves maintaining data availability, data integrity, and data security (Mosley, 2008b). To be most useful internally and to our partners, an always-on situation is ideal. Always-on means the system should be accessible 24 hours a day, 7 days a week, and for an indefinite length of time. Physical data integrity becomes critical in this type of environment. Hardware will degrade with constant use over a relatively short period of time and can ultimately fail. The data management system should be designed to detect hardware failures on the fly and have replication and back-up procedures in the event that a hardware component does fail. Other factors that threaten data integrity include obsolete data formats and data corruption. Datacasting, the ability to convert data from one format to another, is becoming increasingly common. There are many open standards for converting from one data type to another. At the time of data preservation, the author or collector of the data should store the data in the most suitable format to preserve accuracy and integrity. The database manager should maintain the necessary format dependencies on the server and configure the data management system to cast the data to other desired formats. Data corruption can be unintentional or can be the result of a security breach. Data may accidentally be modified by bugs in new code updates or by the result of a malicious act. The data management system needs to be equipped to handle end-to-end checksumming of data bytes to ensure data corruption never occurs as a result of computer error. Additionally, to protect from malicious acts, it is vital to have a user account system that keeps track of who uploads and modifies data. All of these components fall under a traditional information technology audit. Moreover, it takes a dedicated staff and funding to oversee these operations, factors that are often the largest challenge for smaller groups.

Strategies and Lessons Learned

There are obvious challenges to implementing a data management plan and system, but nonetheless, data management is a key aspect in continuing to support the USGS mission. To ensure the success of the project, several strategies are being considered to help overcome challenges. Below are the strategies to be pursued in working toward the project goals, given staff, resource, and funding realities.

Resources and Funding

Leveraging resources has thus far been the greatest factor for success. Often by leveraging resources, it is possible to begin a project with little risk, as initial investment costs in personnel and hardware are minimal. The USGS-wide resource, ScienceBase, has been relied upon to explore development of a data management system without investing heavily up front. By providing an existing infrastructure, ScienceBase has allowed a data management plan to be developed quickly for the Chesapeake Bay studies. A custom subset of the entire ScienceBase data repository was created and used to begin thinking through the stages of the CDI scientific data life cycle model. ScienceBase has provided the opportunity to talk with the research community in a more concrete manner, rather than using hypothetical mock-ups and descriptions. As a result, the use of ScienceBase has provided a better defined set of data management requirements for the future.

Leveraging resources goes hand in hand with avoiding redundancy and has both a short-term and long-term gain for this project. Using the ScienceBase example, avoiding redundancy was a logical choice for the project because limited staff and funds were available to dedicate to building and maintaining a new server. In addition, the intention is to not only avoid creating duplicate data management system frameworks but also to avoid data redundancy. External partners produce and maintain many key datasets used in USGS Chesapeake Bay research, but it would not be practical to maintain those datasets. The goal in future stages of this project is to work more closely with external partners to make their data harvestable by the ScienceBase or any future USGS Chesapeake Bay studies data management system.

Staff and Workflow

There are common methods to help organizations cope with changing staff roles and workflows as a result of technology changes. In general, increasing user involvement helps overcome user resistance (Laudon, 2011). Three key activities that will be implemented to make the personnel transitions more successful are to hold informational meetings about data management, offer employee training on the new data management system, and create incentives. The annual USGS Chesapeake Bay Workshop has provided a good opportunity to reach critical staff in the past, but as data management activities increase, more informational sessions in the form of webinars will need to be scheduled. Training also needs to be provided on use of the data management system to overcome user frustration and increase adoption of the system. Because ScienceBase was used as the initial data management system, USGS-wide webinars can be used as training material and smaller Chesapeake Bay activity training sessions can be coordinated with ScienceBase technical support staff. Both training and informational sessions are activities that can be coordinated within the USGS Chesapeake Bay activities group, but creating workflow incentives is a much more complicated task and will take high-level and consistent adoption across the USGS to create workflow incentives. The USGS CDI community has begun discussing the issues surrounding changing workflow incentives; the Chesapeake Bay studies will continue to be involved and provide useful cases to support those discussions.

References Cited

Data Management Working Group, Best Practices Focus Group, 2012: U.S. Geological Survey web page accessed June 20, 2012, at *https://my.usgs.gov/confluence/display/cdi/DMWG+Best+Practices+Focus+Group*.

Fort Collins Science Center Web Applications Team, 2012a, U.S. Geological Survey community for data integration: Data upload, registry, and access tool: U.S. Geological Survey Fact Sheet 2012–3074, 2 p., accessed September 18, 2012, at *http://pubs.usgs.gov/fs/2012/3074/*.

Fort Collins Science Center Web Applications Team, 2012b, The USGS ScienceBase Catalog—A "Mother Lode" of Science: U.S. Geological Survey Web page accessed June 21, 2012, at *http://www.fort.usgs.gov/WebApps/SciBase.asp*.

Laudon, 2011, Management information systems (12th ed.): Prentice Hall, 640 p.

Mosley, Mark, 2008a, DAMA dictionary of data management (1st ed.): Technics Publications, 148 p.

Mosley, Mark, 2008b, DAMA-DMBOK functional framework, version 3 : DAMA International, 18 p., accessed July 12, 2012, at *http://www.dama.org/files/public/DMBOK/DI_DAMA_DMBOK_en_v3.pdf,*.

Phillips, S.W., 2011, U.S. Geological Survey Science for the Chesapeake Bay Restoration: U.S. Geological Survey Fact Sheet 2010-3081, 2 p., accessed September 18, 2012, at *http://pubs.usgs.gov/fs/2010/3081/*.

Prepared by:
USGS Publishing Network
Raleigh Publishing Service Center
3916 Sunset Ridge Road
Raleigh, NC 27607

For additional information regarding this publication, contact:
Director, EGSC
U.S. Geological Survey
12201 Sunrise Valley Drive, MS-521
Reston, VA 20192
http://egsc.usgs.gov/contactus.html