U.S. Geological Survey Position Paper
Submitted to
Workshop on Research Data Lifecycle Management

June 3, 2011

The U.S. Geological Survey (USGS) is the Nation's and the world's leading natural science and information agency. Its workforce of 10,000 scientists and support staff, distributed in 400 locations, collects and interprets data from tens of thousands of biologic, geologic, and hydrologic sampling sites throughout the Nation. These efforts, combined with extensive remote sensing and modeling capabilities, allow the USGS to map, visualize, and understand Earth processes and changes.

The USGS maintains a broad scope of research activities and long-term data sets relating to earthquakes, tsunamis, and volcanoes, energy and mineral resources, real-time flood data, surface- and ground-water resources, and information critical to dealing with invasive species and ecosystems. Our customer needs and internal business processes are evolving as they relate to the science disciplines we represent. Robust digital descriptions that can be discovered, evaluated, and accessed much more readily in an internet-available form for scientific computing are now required. The dynamic nature of these needs presents a fundamental shift in our historical approaches for curating USGS scientific data content.

The USGS Community for Data Integration (CDI) was established in 2009 to develop and execute a plan for USGS data resources to facilitate discoverability, improving usability for scientific computing, developing consensus on relevant data products, and enabling data integration. The CDI provides a forum to focus on data integration issues, planning, and execution, and to assist in providing Bureau level guidance to implement the USGS Data Integration Strategy. The CDI's responsibilities include:

1. Leading development of the data integration strategy
2. Providing data integration recommendations
3. Promoting data integration Bureau-wide
4. Cross-collaborating with Federal Agencies to refine our data integration efforts

The CDI sponsors a Data Management Working Group that captured a scenario it believes is typical across our science agency:

... the lifecycle of a data set does not end with a given scientist or project. The ability to integrate multiple datasets for analysis and reuse expands the reasons for which a single dataset was originally collected. Data collection and analysis is only part of the foundation of science. Data integration is another key component needed to answer more complicated questions in science. However, before data integration can be undertaken, it requires the data to meet certain standards that define the data lifecycle. There is an underlying assumption in USGS that the majority of data is available and poised for integration. This is simply not the case for most data as in many offices and programs,

scientists and managers lack guidelines and standards to help ensure that relevant and critical documentation is collected before, during, and after data is collected. Scientists spend needless time and money reproducing data sets that have already been collected, because they are unable to locate pre-existing collections. Historical analyses are unable to be conducted because relevant data sets are missing necessary contextual information. In the current business model, it is difficult to find data within the Survey, much less to access and understand it.

The CDI Data Management Best Practices sub-team is working to locate or develop a hybrid data lifecycle model that best matches how scientific data is created, used, preserved, and reused within and beyond the USGS. The resulting model will be used to help align data management best practices. The group developed guiding principles to include *simplicity* to appeal to scientists who will need to understand it, a model that is fairly *intuitive* on its own, and where possible, identify a *separation of responsibilities* between scientists and support staff in order to lessen the load on the scientists.

Over the last seven months the team has collected 20 data lifecycle models obtained through literature searches. The sources include, the Digital Curation Centre based in Scotland, the Federal Geographic Data Committee, the University of Oxford, the National Oceanic and Atmospheric Administration, the Environmental Protection Agency, the Office of Science and Technology Policy's Interagency Working Group on Digital Data, the Bureau of Land Management, and original USGS contributed models. The applicability of these models to how our science data are created, maintained, and archived is currently being evaluated. The model selected, or hybrid developed, will become the foundational element from which we will attach our own best practices such as scientific records appraisal processes and electronic records management procedures. It is envisioned that the model will need to be comprehensive and at the same time not overly complex so that it can be easily implemented. The best practices may appear at points within the data lifecycle as specific guidance, tools, standards, templates, or sources of assistance. Many of these best practices are already in use within the Agency, but in an inconsistent manner. A data lifecycle model will be the tool to organize and deploy these practices effectively and provide a foundation upon which future data management and integration activities can be built upon.

The CDI is endorsed by senior USGS management and driven by a passion to preserve and make accessible the science data our agency creates. The Workshop on Research Data Lifecycle Management would be an opportunity for USGS to contribute and share our work to date and how we plan to utilize such a model. Allowing USGS participation gives the greater scientific data community an important voice with which to influence our own plans while helping us better serve societal needs. The timing of the Workshop coincides well with our efforts in examining models to locate or develop the most appropriate one for our own use.

USGS participation would also contribute directly to any discussion on the assessment and selection of research data. Over five years of experience can be shared involving a formal scientific records appraisal process that is recognized as a National Archives and

Records Administration best practice. This process relies heavily upon stakeholder involvement consisting of an archivist, relevant scientists, and programmatic management. Several steps are formally documented so that decisions can be justified and if need be, reappraisals can occur at later dates involving much less time and energy.

As pointed out earlier, the USGS is in the midst of gathering lifecycle model use and implementation ideas. Work can also be shared involving metadata and data discoverability/accessibility. Standards efforts and recent mandates to share all data openly and at no cost have provided us with incentives to document thoroughly and make the discovery, access and usability research steps easier thereby reducing duplication of effort and data. The end result is better and more accessible science preserved for the long-term.

In summary, the USGS would hope to both actively contribute to and take from the Workshop on Research Data Lifecycle Management. The timing of the Workshop coincides well with our efforts in examining models to locate and develop the most appropriate one for own agency use.