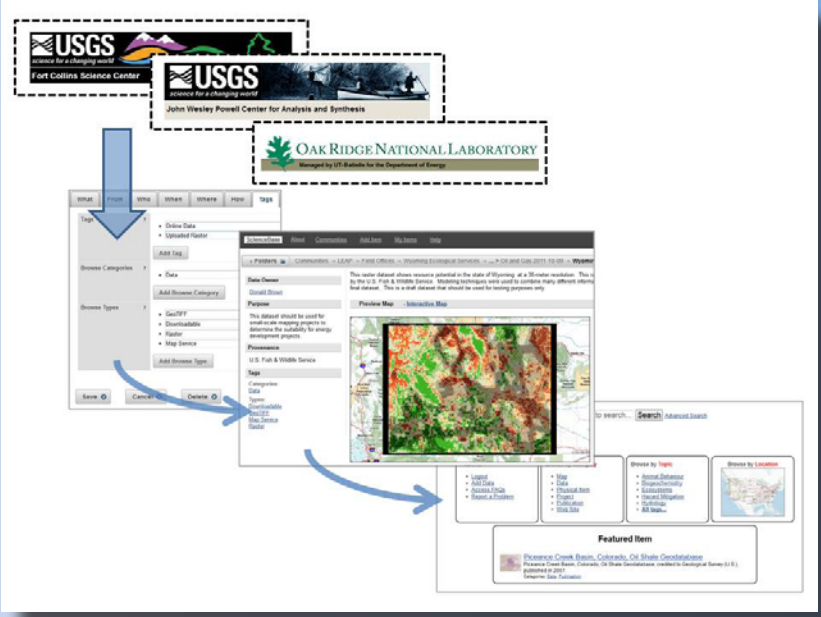
U.S. Geological Survey Community for Data Integration: Data Upload, Registry, and Access Tool

As a leading science and information agency and in fulfillment of its mission to provide reliable scientific information to describe and understand the Earth, the U.S. Geological Survey (USGS) ensures that all scientific data are effectively hosted, adequately described, and appropriately accessible to scientists, collaborators, and the general public. To succeed in this task, the USGS established the Community for Data Integration (CDI) to address data and information management issues affecting the proficiency of earth science research. Through the CDI, the USGS is providing data and metadata management tools, cyber infrastructure, collaboration tools, and training in support of scientists and technology specialists throughout the project life cycle.

The Uploader allows scientists with limited data management resources to address many of the key aspects of the data life cycle.



Data from individual scientists and laboratories, often on secure internal networks, can be quickly uploaded and made available to all ScienceBase Catalog users worldwide.

What It Is

The Data Upload, Registry, and Access Tool, or "Uploader," involves much more than data uploading. The Uploader provides a project-oriented workflow that consolidates two crucial aspects of data management: (1) uploading and storage of data into an accessible repository and (2) capture of adequate metadata into a searchable catalog. The Uploader performs these functions to enable scientists to find, access, and visualize project data through interactive maps.

The Community for Data Integration defined a number of project goals or "user stories" that demonstrate situations in which the CDI and the products it generates (such as the Uploader) can be put to practical use in support of USGS scientists' work, as follows:

- As a USGS scientist, I have data that I have created or enhanced from other data, and I need a way to post these data online, document them so that others can understand them, and share them with colleagues before they are published.
- As a USGS scientist, I need to be able to make project data available publicly as part of the publication process after my findings based on these data are published in a journal or USGS report. Ultimately, I would like to "link" or relate these datasets to my resulting publication.
- As a USGS data owner, I want to describe how my data should be presented, ensure that any data visualization is consistent with my data descriptions, and make specific data available for download or through Web services.
- As a USGS scientist, I want to discover and review project datasets that pertain to my current/proposed study and easily include relevant pieces or subsets into my research.

The Uploader addresses all of these user stories, providing a relatively easy way for scientists to upload and capture data, to control access to those data, and to distribute those data directly to users or to data portals.

How It Works

The Uploader is a Java-based online tool that combines a number of open-source applications and libraries inside a custom framework. The tool is built around well-supported industry standards, including a representational state transform (REST) based architecture, open source relational database storage of geospatial data, and an open source document-oriented approach to metadata capture. It significantly borrows concepts developed through a number of other data repository and visualization efforts, including the Unidata RAMADDA (University Corporation for Atmospheric Research, 2011) project.

The application uses Apache Tika (The Apache Software Foundation, 2012a) to infer metadata from ingested files, the MongoDB

(10gen, Inc., 2012) for cloud-compatible data management, Apache Lucene (The Apache Software Foundation, 2012b) for search, and a variety of other third-party software for data access, visualization, and management (U.S. Geological Survey, 2012). The Uploader provides a custom security layer and cross-utility interfaces to tie these diverse projects into a seamless view for the user.

Data management also requires provisions for data protection and preservation. To address these requirements and ensure a stable long-term repository for these datasets, the Uploader is integrated into the USGS ScienceBase platform (http://www.sciencebase.gov/catalog), thereby allowing the integrated Uploader to take advantage of the hosting, architectural, replication, archive, network, and security capabilities provided by ScienceBase.

The Uploader provides an easy way for scientists to upload and capture data, control access to those data, and to distribute those data directly to users or to data portals.

The Uploader provides an easy way for scientists to upload and capture data, control access to those data, and to distribute those data directly to users or to data portals.

Who Can Use It

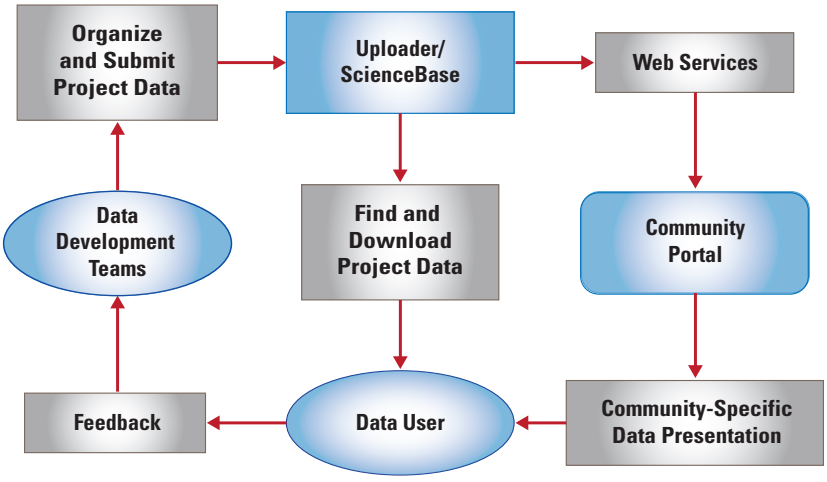
Scientists in the USGS and partner organizations can use the Uploader, which is available from within ScienceBase and accessed through custom community portals, such as the Landscape Conservation Management and Analysis Portal (LC MAP), the Landscape Energy Action Plan (LEAP), and the Powell Center Extraction, Visualization, and Analysis project. Those wishing to include the Uploader and ScienceBase into their science project workflow should contact the FORT Web Applications Team at the link provided.

The Uploader allows scientists with limited data management resources to address many of the key aspects of the data lifecycle: the ability to protect, preserve, publish and share data. By implementing this application inside ScienceBase, scientists also can take advantage of other collaboration capabilities provided by the ScienceBase platform.

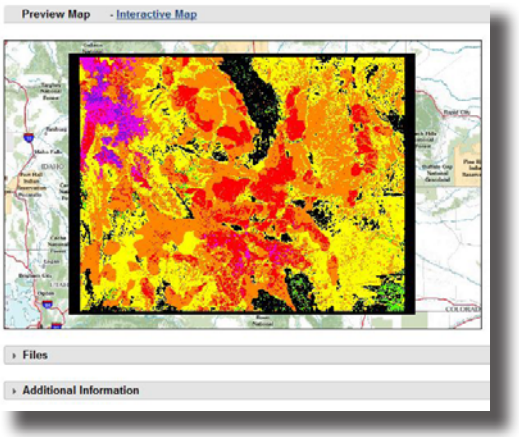
Benefits to USGS Science

Just as each community can maintain and share project datasets, references, metadata, and reports among their members, communities can share items with other communities. This culture of data sharing is the first—and perhaps most crucial—step to effective data integration. For example, several of the U.S. Fish and Wildlife Service Landscape Conservation Cooperatives (LCCs) have implemented this integration model (as shown below) into their data management strategy. Western LCCs maintain their own items in LCC-specific catalogs, and LCC data stewards can then share approved items with the LCCs at large through LC MAP. In this way, all LCCs can learn about and take advantage of the work conducted by an individual LCC.

The Uploader project is an example of the identification of business requirements, leveraging of existing software applications, and collaboration by partners to create a successful tool. Use of the Uploader tool continues to grow, resulting in more data available for broader integration with current and future projects, providing as yet untold additive contributions to science and our community knowledge for future USGS research.



The Uploader is part of a continuous workflow that promotes data management best practices for the science team and data sharing best practices for all users.



The Uploader enables scientists to find, access, and visualize project data through interactive maps.

References Cited

10gen, Inc., 2012, MongoDB project documentation: MongoDB, accessed June 20, 2012, at http://www.mongodb.org/.

The Apache Software Foundation, 2012a, Apache Tika project: Apache Tika, accessed June 20,2012, at http://tika.apache.org/.

The Apache Software Foundation, 2012b, Apache Lucene project: Apache Lucene, accessed June 20, 2012, at http://lucene.apache.org/.

University Corporation for Atmospheric Research, 2011, Unidata RAMADDA documentation: Unidata Program Center, accessed June 20, 2012, at http://www.unidata.ucar.edu/software/ramadda/.

U.S. Geological Survey, 2012, ScienceBase documentation: ScienceBase, accessed June 20, 2012, at http://www.sciencebase.gov/.

Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Authored by the *Fort Collins Science Center Web Applications Team.*

For more information, contact:

Director, Fort Collins Science Center
2150 Centre Ave., Building C
Fort Collins, CO 80526
http://www.fort.usgs.gov

Community for Data Integration
Associate Director for Core Science Systems
U.S. Geological Survey
108 National Center, 12201 Sunrise Valley Drive
Reston, VA 20192
703-648-5748
http://www.usgs.gov/core_science_systems/