ScienceBase: data and information management for scientists

Sky Bristol, Natalie Latysh, and Tim Kern

www.sciencebase.gov

Abstract

The U.S. Geological Survey (USGS) is working to enhance and expand information sharing and preservation by developing ScienceBase, a collaborative scientific data and information management framework used by scientists and data practitioners. ScienceBase is a long-term, strategic commitment to solving some of the biggest problems in managing, accessing, and using scientific data and information. The ScienceBase effort is both one of tools and technology as well as the adoption and augmentation of methods and practices at an enterprise scale and within small science teams.

What's in ScienceBase

"Getting to the data is good, but I also need to track back to the scientific manuscripts where the data are described and analyzed; and then I need to be able to understand who worked on the data and what projects they came from."

Science teams need a whole collection of information sources brought together in one place to serve as a foundation for their work. They need to start with what ScienceBase already knows that might be important and then add to it with new records and additional information on the records already in ScienceBase. We started partially with the concept that "Everything is Miscellaneous"* and that one person's publication is another person's data, but we did need to come up with a high level classification that includes the 7 different types of items listed here.

* everythingismiscellaneous.com

Ways of organizing

Keywords are an important way of organizing information, but they are also a bit problematic in a larger scale cataloging effort. ScienceBase does a lot of harvesting of information items from scientific publication sources to other major data catalogs. Many of these come with their own keywords, tags, or other ways of describing records, and not all of them specify some specific controlled source. Right now, ScienceBase lumps all of these together and presents them to users as clickable tags that run further searches to find other like items. On the horizon, ScienceBase is building a more elegant vocabulary registry that will work to make sense between the different types, sources, and granularity of the terms flowing into the system.

People are important

Information about authors and other types of contacts for everything from publications to data sets to projects are a really important connecting factor. Metadata doesn't always answer all our questions about how to use a particular data set, and we often really care about getting a hold of other work from a particular scientist. ScienceBase puts a fair bit of energy into using people and organizations as an integration point across disparate data sets. The ScienceBase Directory is an element to the framework that works to store all the known information about people that is augmented over time as we learn more and work to actively maintain quality information.

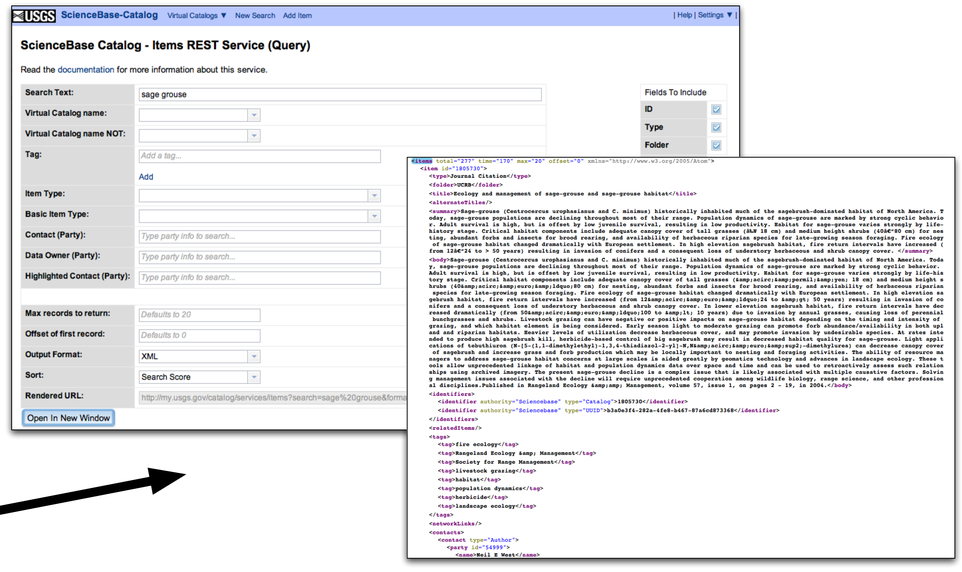
Supporting the unanticipated good uses of ScienceBase

Even through we need to provide a solid user interface to search ScienceBase, manage information in ScienceBase, upload data, and perform other functions, we can't anticipate all the good uses that people may make of the information we harvest and make more accessible or that science teams create as part of their work. ScienceBase provides a variety of web services that return the same results seen in the search interface via XML and other formats. The simplest is an ATOM feed that packages most of the information in a simple format. Services can be plugged in to web sites and other tools for a variety of uses. We are experimenting with the "importXML" functions in various spreadsheet platforms to show how ScienceBase information can be retrieved into a completely different tabular format and used to build a variety of reports and other presentations. The ATOM response from the ScienceBase REST service is currently being used with web sites built with Drupal* and and the "Feed Aggregator" module to show syndicated content from ScienceBase in a completely different layout.

* drupal.org

Searching and Browsing

Many users of ScienceBase asked for a "Google-like" experience of searching. We're still trying to achieve that level of sophistication with a different set of search parameterization and weighting (geared toward scientists), but we are also working to make sure that ScienceBase information is exposed in an organized way for all the big search engines to pick it up and provide other avenues for discovery (e.g., site indexes to individual item pages, exposure of records as KML, and seeding of other indexes). The core ScienceBase interface provides simple and advanced searches, ways of filtering results, and different ways to browse what's in ScienceBase.



Key Features of ScienceBase

1. Data uploading, documentation, and sharing (selectively or publicly as appropriate to the content)
2. Metadata and data services to enable other applications and web sites use the information in different ways.
3. Harvesting engines that go out and bring in important information resources from big catalogs to small imports or selective records.
4. Ability for science and management teams to find and organize information resources important to their work and to provide additional attributes on existing records to help add meaning to ScienceBase
5. ScienceBase data managers and consultants who work to improve information discovery by making connections across sources using key elements such as contacts and help science teams in their endeavors

Place is important

A lot of what we do in earth science is spatially referenced in some fashion. ScienceBase uses the concept of a geospatial footprint to help put things on a map. A footprint can be anything from a geotag that gives us some idea of where to show an item and return it in results to representational point, a bounding box, or a more complex geospatial feature. Any item can have any combination of these that are used in different ways depending on the question being asked, but it is also important to note that not all items are at least explicitly able to be put on a map. Many footprints in ScienceBase come directly from a source metadata record. ScienceBase also provides a component we call Footprint Studio where users who are managing records right in the system can find features on a map such as a hydrologic unit, draw features on the map, or upload a shapefile containing a more complex geometry. Future research in this area is focusing on more advanced ways of exposing spatial intersections between disparately discovered items, helping scientists to make more rapid connections based on the underlying features of an area on a map (e.g., similarity between items based on hydrologic or geochemical landscape setting).

Context is important

Context is an important concept in ScienceBase. It factors in when we work on harvesting important information sources for the things ScienceBase users want to put together. Many information sources and even large catalogs and aggregators place their contents into a particular context and rely on the fact that you are in that context to know a little bit about what the items mean. For instance, the title listed in the metadata for a web service provided by an ArcGIS Server might assume that you are looking at that service in the context of a mapping application that tells you what the map is all about. Harvesting and presenting just the information about that service to someone through ScienceBase without concatenating a little more information from a parent container or some other source might not tell someone enough to know whether or not the service might be meaningful.

Context is also important for the work that a given science team may do over the course of a project or some ongoing work. ScienceBase uses the idea of context to give science teams a virtual catalog for their work. Within a virtual catalog, science teams can bring together resources they find that are already in ScienceBase and upload and create new records. The context gives some additional meaning to the records being assembled and can be used to present those records through services and other means. Future work in ScienceBase is looking to align the context concept with the Open Geospatial Consortium Web Map Context standard such that contexts and their associated information can be used in a wider variety of ways through ScienceBase services.

Like to get involved?

Visit us at http://www.sciencebase.gov/.

We are currently in the process of modularizing several key elements in the ScienceBase architecture such that they can be released as open source products, both as part of the overall ScienceBase body of work and through avenues such as plugins for various frameworks (e.g., Grails*, Drupal**). We hope to encourage involvement in building and enhancing the suite of tools that makeup ScienceBase with developers working within compatible frameworks.

We also welcome scrutiny and feedback on the public interfaces to ScienceBase. Check things out and use the feedback link to give us your ideas.

ScienceBase 2.0 is on the horizon, and the biggest part of that work revolves around a major retuning of the ScienceBase Item information model and alignment of the model with applicable standards from ISO and FGDC metadata to the map service standards being used for uploaded and served data sets. We welcome review of these concepts and the developing information model by visiting the ScienceBase web site and looking for the "ScienceBase Brain."

Key partnerships in the ScienceBase endeavor include the U.S. Geoscience Information Network (usgin.org) and DataONE (dataone.org).

* grails.org
** drupal.org