ScienceBase: a big ol' scientific database

Sky Bristol, Tim Kern, Natalie Latysh, Steve Tekell, David Mack, Jeff Allen, Dell Long & Robb Prescott

www.sciencebase.gov

Abstract

ScienceBase is a U.S. Geological Survey (USGS) effort to reflect in a database all we know about the complex earth system. We admit that it is a challenging vision but we think it is achievable in an agile, incremental way. ScienceBase started off as the Scientific Data Catalog and later the Comprehensive Science Catalog. This first generation of the concept was "yet another metadata catalog," and cataloging of resources is still something that ScienceBase has at its core. The second generation, where we attached the moniker of ScienceBase, has added a data repository capability and started addressing the long tail of dark data in USGS. The third and fourth generations of ScienceBase will take us into the territory of data integration and proactive analytics, respectively. This poster presents key concepts of the current ScienceBase 2.0 architecture.

ScienceBase Items and Collections

ScienceBase is organized around the concept of collections of items, where every record of any type is an item (similar to OWL:thing or RDF:resource). Item information is contained in a document stored as JSON (or BSON in MongoDB), and that native format for items is what drives API functionality. Items can be translated into a variety of other formats (e.g., ISO19115 XML) through the sbTransformer. Some item information is generated on the fly or through caching operations via Map/Reduce techniques. Items are organized into the following types of collections.

- Community Collections - items generated through the interactions of a particular community of use through file uploads, web forms, and other API interactions
- Harvested Collections - items of interest to ScienceBase harvested from another source such as a catalog service or web accessible folder (ScienceBase is generally not considered authoritative for harvested items but may be a best available source in some cases.)
- Native Collections - items that are integrated into and provided by ScienceBase as their new native home

ScienceBase and Files

ScienceBase allows for any file type to be uploaded and "attached" to a ScienceBase item. We have a practical limit of around a 2 GB file upload over HTTPS and are working on a solution for larger file transfer. Several types of files have special handlers that are triggered by their presence in ScienceBase.

- shapefile - WMS, WFS, and KML service endpoints via GeoServer
- GeoTIFF - WMS, WCS, and KML service endpoints via GeoServer
- Esri Service Definitions - ArcGIS REST service via ArcGIS Server
- image/ - all image files can be downscaled and resized on demand through the API; a pre-formatted medium-resolution image is cached for previews
- Excel - spreadsheet files can be processed to create new child items in the hierarchy via a JSON configuration file
- EndNote/MODS XML - recognized as citation formats and can be processed to create child citation items via sbTransformer
- NGGDPP - specialized XML and CSV formats for geoscience samples and other physical artifacts
- NetCDF - (in R&D) processed behind THREDDS data server for OpenDAP services

ScienceBase and Standards

One of the primary cataloging functions of ScienceBase is to serve as a layer of abstraction between various metadata and data standards and multiple, diverse inputs and outputs. Standards are consulted each time new data modeling occurs to determine appropriate mappings and constraints, which are then encoded into the sbTransformer for both data input and output.

- FGDC XML - metadata records can be input and output; not all FGDC elements are currently mapped into the ScienceBase Item model
- ISO19115 XML - items can be output as ISO19115-1 metadata; work currently ongoing to incorporate ISO19115-2 elements
- EndNote/MODS - input XML formats
- ATOM/OpenSearch - ATOM-formatted search results with OpenSearch compliance; alternate format for single item output; ATOM harvester for serialized item input
- OGC-CSW - CSW search capability built into the ScienceBase API
- EML - (in R&D) input ecological metadata format

ScienceBase Application Programming Interface

Retrieval Services

The ScienceBase API can be interfaced with in many ways to generate creative applications. Foundational components include catalog services, various metadata transformations, faceted searching, and inherent map services at multiple levels. The RESTful architecture means that all data/metadata retrieval functionality can be saved in a simple URL.

Repository Services

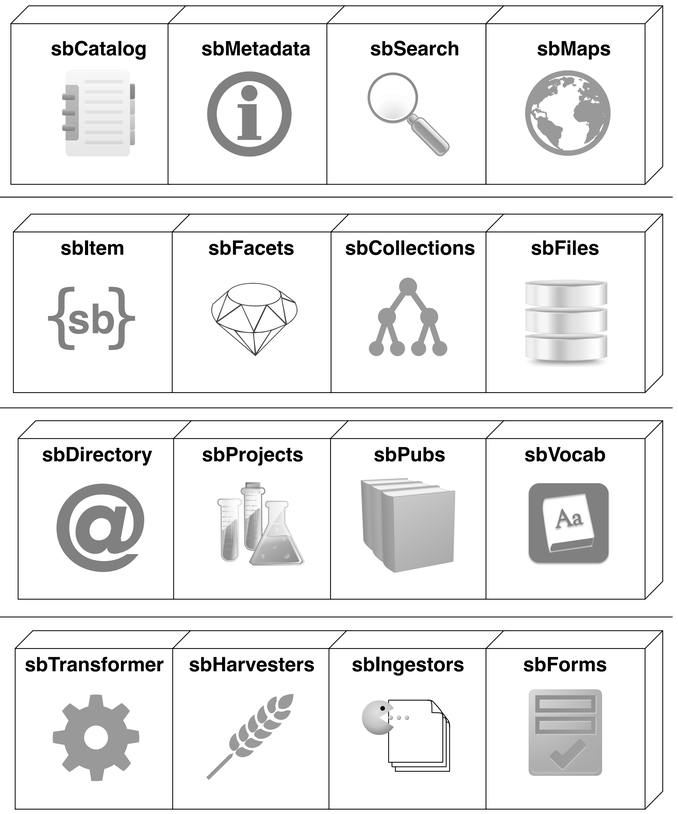
The ScienceBase data and metadata repository consists of items with simple core attributes, a growing number of facets containing extended information and functionality, an inherent hierarchy that supports permissions and methods of interacting with items in collections, and a robust file store containing various kinds of file "attachments".

Master Data Services

ScienceBase provides a number of corporate master data, integrated together into the ScienceBase item structure as collections. These include a directory of known people and organizations, a collection of active and completed science projects, the complete collection of USGS publications, and an evolving vocabulary service.

Data Input Services

Items come into ScienceBase in a variety of ways that nearly always involve some form of transformation from source material to the ScienceBase item model. Methods include harvesters that go out and get items, ingestion engines that process source files of different kinds, various form applications, and methods of direct interaction with the API to POST/PUT/DELETE items.



Overview of ScienceBase architectural components. Refer to additional details in surrounding sections.

Current Technologies

ScienceBase leverages many different open source projects and is itself released as a public domain source package. The one exception to the commitment to open source is the use of the ArcGIS Server platform for Esri-specific file formats and functionality required within the USGS community. ScienceBase has been through two distinct generations of technology with a complete refresh to the tech stack. Future iterations will likely repeat this trend as new capabilities become needed and apparent.

- MongoDB - NoSQL database, document store
- ElasticSearch - text and spatial indexing and search
- REST API - built with myriad Java libraries and custom code
- Grails - ScienceBase management and flagship search UIs
- GeoServer/GeoTools - OGC spatial services engine
- Esri ArcGIS Server - ArcGIS REST services for Esri-specific data formats, images services, and geoprocessing services; live spatial data versioned editing

ScienceBase Services

As of ScienceBase 2.0, the platform is completely built around the idea of web services and a robust API being the primary development and engineering focus. All user interfaces are built on the API as a way of "eating our own dog food" and setting examples for the possible. Underlying components such as MongoDB and ElasticSearch provide their own APIs that ScienceBase leverages along with other functionality with the following custom-built API elements:

- Item services - key to all interactions with the ScienceBase Item structure allowing full CRUD operations on the data system
- Item hierarchy services - items arranged hierarchically become collections that can be interacted with through the API; geospatially referenced child items generate WMS, WFS, and KML at the parent via GeoServer
- Map services - sbMaps provides geospatial services for spatial data associated with items, OGC services, Esri services, static map PNGs
- Directory services - sbDirectory contains all people and organizations ScienceBase becomes aware of; provided as a directory service for direct interaction and through the overall data platform
- Vocabulary services - beta version of the vocab service functions mainly as a code list system; full blown registry of controlled vocabularies and ontology services in R&D
- File-driven services - ref. ScienceBase and Files

API-driven Applications

ScienceBase follows an API first principle of building all functionality on top of a robust Application Programming Interface. We are beginning to see more and more "powered by ScienceBase" applications built across the USGS and with public and private partners.

- Drupal - modules and templates for query, reporting, mapping, and other features
- myUSGS - ScienceBase extentensions for Atlassian - Confluence, JIRA, Stash
- DEPTH - custom JQuery app for managing project records and generating reports
- Data Basin - shared ArcGIS services and leveraged ScienceBase upload and repository capability
- GeoDataPortal - geo-processing services (OGC-WPS) for downscaling climate models; workflows reposited in ScienceBase; direct API to API integration
- RFP Manager - custom Grails app for managing proposal processes and data management planning

Guiding Principles

"Everything is miscellaneous"

The "unsolvable" problems are the most interesting

Technology is ephemeral

API first

July 2013