

The Open Data Repository's Data Publisher

N. Stone¹, B. Lafuente², R. T. Downs², D. Blake³, T. Bristow³, M. Fonda³ and A. Pires²



IN21E: Agile Curation, Data Access and Infrastructure, and Data Layers:
Maximizing the Value of Research Data II

December 15, 2015

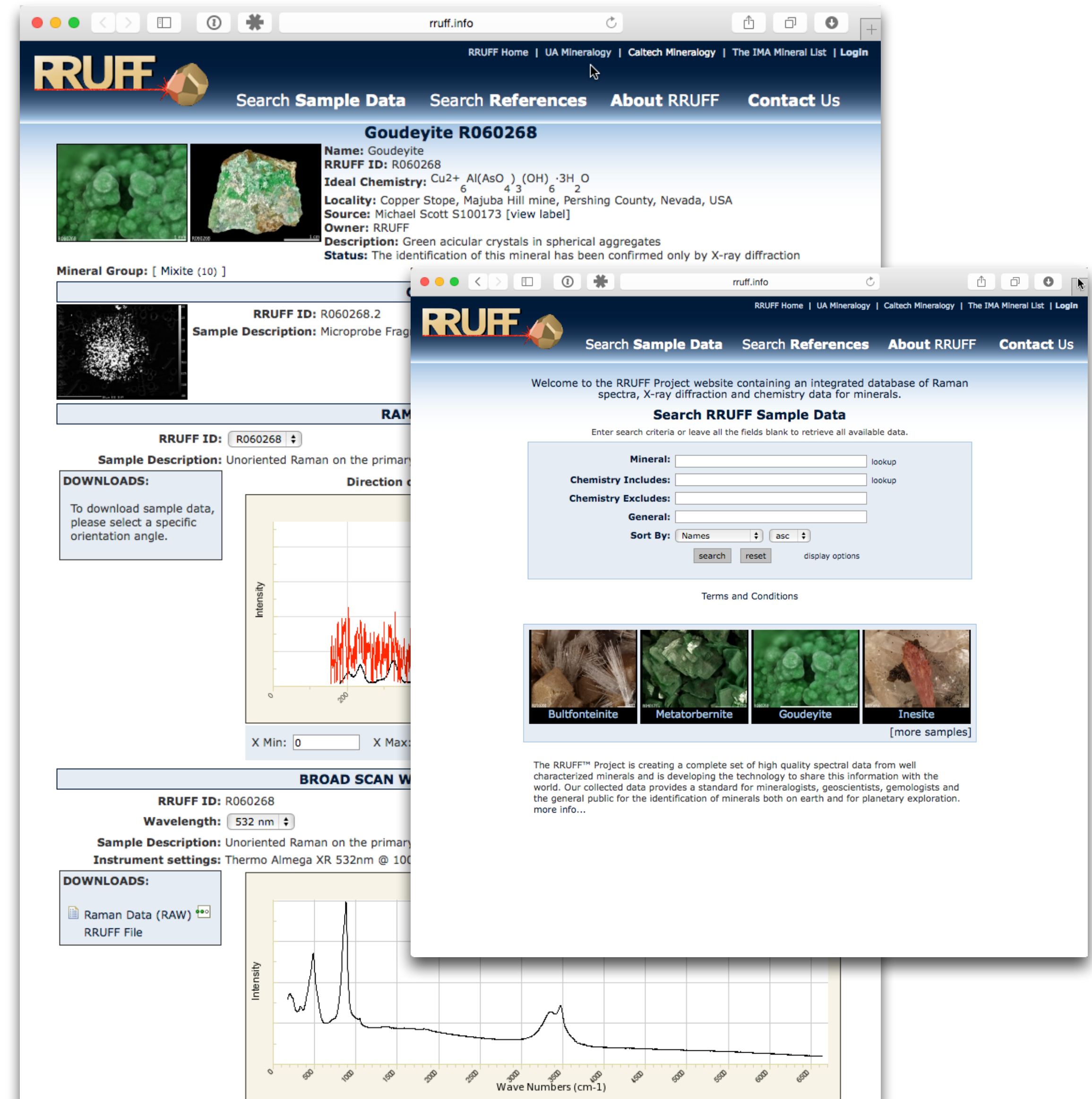
We gratefully acknowledge the support for this study by the Science-Enabling Research Activity (SERA), and NASA NNX11AP82A, Mars Science Laboratory Investigations and University of Arizona Geosciences.

Nate Stone
Open Data Repository
nate.stone@opendatarepository.org
<http://www.opendatarepository.org>

¹Open Data Repository, Gray, ME, USA
²Geosciences, University of Arizona, Tucson, AZ, USA,
³NASA Ames Research Center, Mountain View, CA, USA.

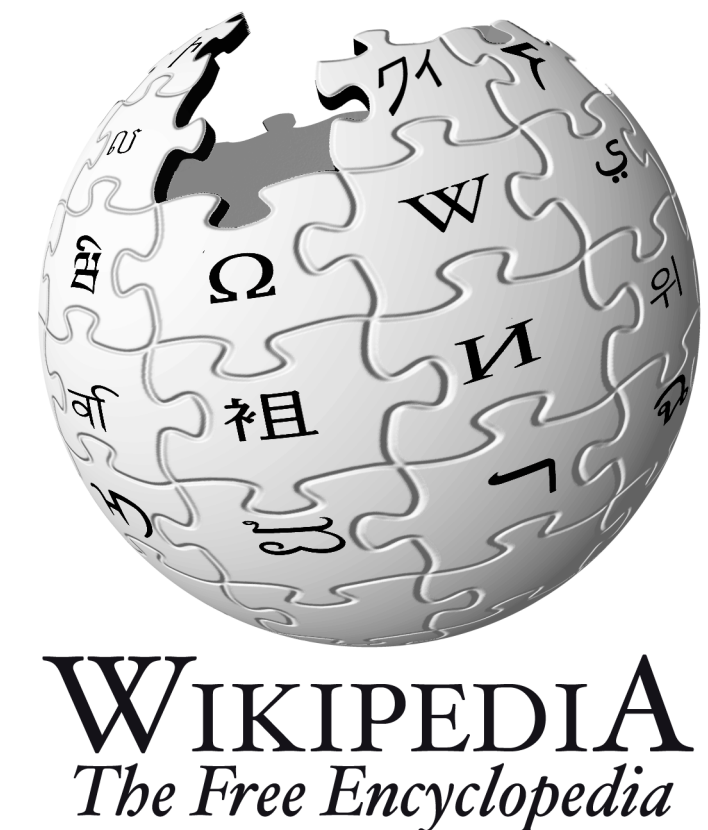
History: The RRUFF Project

- The RRUFF Project houses spectra and information about nearly every naturally occurring mineral on earth.
- Over 400MM hits to the database since its inception in 2004.
- Web-based interface. Easy to search using Google or its internal search.
- Easy to connect with - many applications directly download spectra and data for use in real time.
- Custom programmed for this data. Expensive to modify what it can store.
- Many scientists have requested a “copy” to store their data.
- Can we abstract it to store and plot nearly any data?



Data Management System

- Bob Downs (University of Arizona) and David Blake (NASA Ames) conceived the idea of abstracting the RRUFF database so that other researchers could organize and share their research data.
- Inspired by Wordpress - shown that simple content publication tools result in an explosion of new material.
- Wikipedia and the WikiMedia Foundation - our goal is to create the software and provide it free so that anyone can create a database and put it online.
- Tailored for publishing and archiving research data.



Project Plan

- Project broken into two phases:
 - Phase 1: Use-case pilot
 - Determine if we can make a tool that fits into a researcher's workflow rather than an additional task that comes when the work is complete.
 - Currently in progress
 - Phase 2: Semantic web integration/General release
 - Develop tools to map data to ontologies, allow ontology-based data creation, RDFa, etc.
 - Citation and snapshot-based archiving system.
 - Focus of Q2-Q4 2016

Market and Objective

- Target the long tail of research data that fall outside of larger initiatives.
 - Accessible to individual researchers or small teams.
 - Runs on a standard LAMP (Linux, Apache, MySQL, PHP) server.
 - Hosted service or host themselves to provide options when there are ownership issues.
 - Doesn't require a specialist or significant training to run.
- Determine what types of data we would be able to support with a simple, web-based tool (spectra, sample collection data, site data, etc.)
- Help researchers organize their data from collection through analysis and finally to preservation.
- Provide a complete audit trail and provenance chain: raw data through to presented analysis.
- Facilitate sharing and re-use by easing mapping to metadata standards, ontologies and providing semantic searches across databases. (Phase 2)
- Archiving of data snapshots/releases within the system and publishing of snapshots to third party repositories. (Phase 2)

ODR Data Publisher: Phase 1 State

- Concept is to create a simple tool allowing researchers to publish and collaborate for the full life cycle of their research.
- Digital laboratory notebook.
- Collaborative posting of data sets and online discussion before final publication.
- Selective/embargoed publication of data sets.

Created by: Alex Pires on 2014-06-11 11:06:03 (UTC-5) Last Modified by: nate@opendatarepository.org on 2015-01-20 04:01:42 (UTC-5)

[Edit DataRecord](#) [Previous](#) [Return to Search Results](#) [Next](#) **Browse Search Results** **Record 2 of 21784**

Search Result » UAMM-Samples

Image

UAMM ID
2

Name
Linarite

Ideal Chemistry
PbCu2+(SO4)(OH)2

Mineral variety
Associated Minerals
Malachite, Azurite

Class of object
☒ Mineral

Quality
☒ Display

Type sample

No Images Found

Locality
Mammoth Mine, Mammoth District, Tiger, Pinal County, Arizona, USA

Museum Label Locality

Revised locality when required.

Locality Notes

[Dashboard](#) [Search](#) [nate@opendatarepository.org](#) [Logout](#)

Display Records - UAMM-Samples

10 records found. Show: 10 25 50 100 Records per page

Page 1 of 1 1

Image

UAMM ID
200

Name
Gold

Locality
Deo Gratias Mine, Cienega District, Sonora, Mexico

Acquisition Source
Fletcher Haskell

Weight (g)
62

Description
Placer Gold in Quartz.

Image

UAMM ID
201

Name
Gold

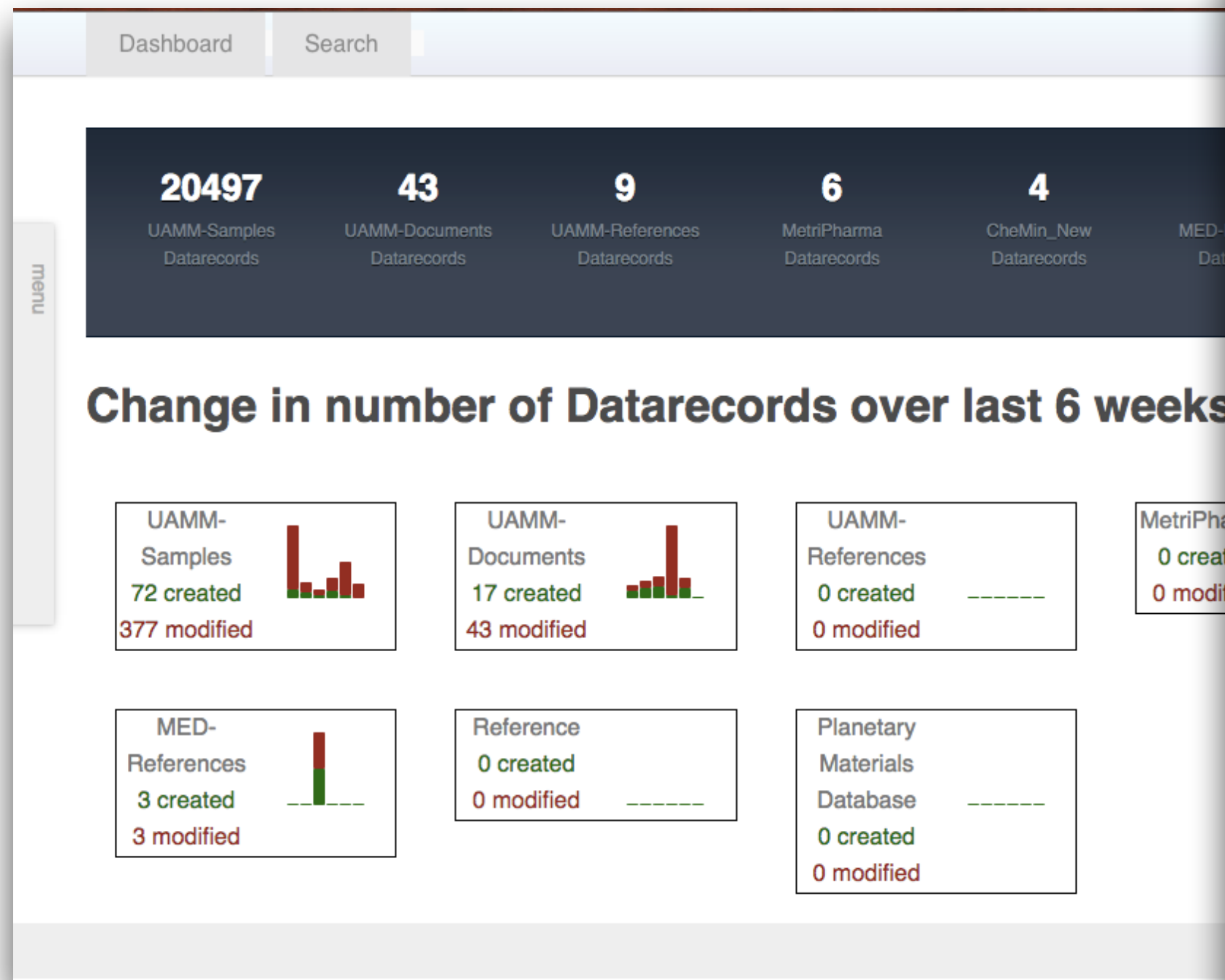
Locality
Deo Gratias Mine, Cienega District, Sonora, Mexico

Acquisition Source
Fletcher Haskell

Weight (g)
3

Description

- Form designer allows quick creation of data structures for searching and display.
- Interface allows a basic web user to create a database and begin uploading data.
- Plugin architecture allows rapid development of customizations such as complex graphing capability, tabular data, etc.
- Fully encrypted file management system with public/private settings for all data.
- Dashboard to view data stats and database activity.
- Linking of databases to allow re-use of data across databases..



This screenshot shows a detailed view of a data record for a CheMin analysis. The page includes a top navigation bar with 'Dashboard', 'Search', and user information (Nate Stone, Logout). Below the navigation bar, there are tabs for 'Edit DataRecord', 'Previous', 'Return to Search Results', and 'Next'. The main content area is titled 'Search Result » CheMin' and contains several sections:

- Gallery of images**: A thumbnail image of a rock sample.
- Sample/calibration name**: Rocknest (4th Scoop)
- Mars Area Location**: From Bradbury Landing to Glenagly Intrigue
- Soil's**: 74
- Aprox. site coordinates**: 4.59°S 137.44°E
- Analysis number**: 1
- Type of analysis**: Scoop
- Description**: A paragraph describing the analysis process and results.
- XRD**: A section for X-ray diffraction data, including a '2D XRD pattern' and a 'List of frames and soils'.
- XRF**: A section for X-ray fluorescence data, including a 'Frame Graph File' and a 'XRF data files Charts'.
- Mineral Abundances**: A section for mineral abundance data, including a 'PDB - MIN file' and a 'Mineral abundance data' table.

The page also includes a footer with 'Open Data' and 'About' links.

Complete Record View

Available Record Features:

- Nested children (n-levels deep)
- Multiple groups of fields each with individual display permissions
- Plugin system for rendering graphs or custom data types.
- Images: jpg, png, gif, tiff, svg, and more.
- File upload/download
- Descriptive text in Markdown
- Public/private permissions on all objects
- Import/Export to XML or CSV
- Link to other records and other databases

DashboardSearch

nate@opendatarepository.orgLogout

Created by: Barbara Lafuente on 2014-12-09 01:12:30 (UTC-5) Last Modified by: Barbara Lafuente on 2015-11-05 02:11:06 (UTC-5)

Edit DataRecord

PreviousReturn to Search ResultsNext

Browse Search ResultsRecord 4 of 12

Search Result » CheMin

Images



DESCRIPTION

Analysis ID

John Klein

Type of analysis

Drill

Sol/s

182

CheMin cell

13b mylar

Description

The John Klein drill sample was collected from the Sheepbed mudstone member of the sedimentary Yellowknife Bay formation, which is interpreted as a shallow lacustrine deposit.

LOCATION

Mars Area Location

Yellowknife Bay

Site position

6_0

Distance to Bradbury (m)

444.81

Latitude

-4.589485

Longitude

137.449129

Elevation (m)

-4519.5

XRD-RDA

sol0196 to sol0269

sol0196 to sol0273

sol0473 to sol0488

XRD Analysis name

sol0196 to sol0269

XRD Product_ID

CMA_414856883RDA01960060000CH12240P1

XRD Start_Time

2013-02-23T06:53:49.000

XRD Analysis Description

This table contains diffraction-all K-alpha diffraction data for the the John Klein drill sample, analyzed in CheMin cell number 13b (Mylar window). The table represents results from sequences uploaded from sol00196 to sol00269, including 9,540 individual 10-second frames in 53 minor frames of 180 individual 10-second frames. This is a preliminary analysis; additional analyses were accumulated after sol 269 and these will be added to the overall analysis in the next data delivery. CCD temperatures during data collection were ~50 degrees centigrade. Column 1 of the table lists 2-theta from 3.00 to 51.95 degrees cobalt K-alpha, in increments of 0.05 degrees (980 entries). Column 2 lists the intensity of the diffraction for each 2-theta value in column 1.

RDA - PDS label file

XY XRD pattern

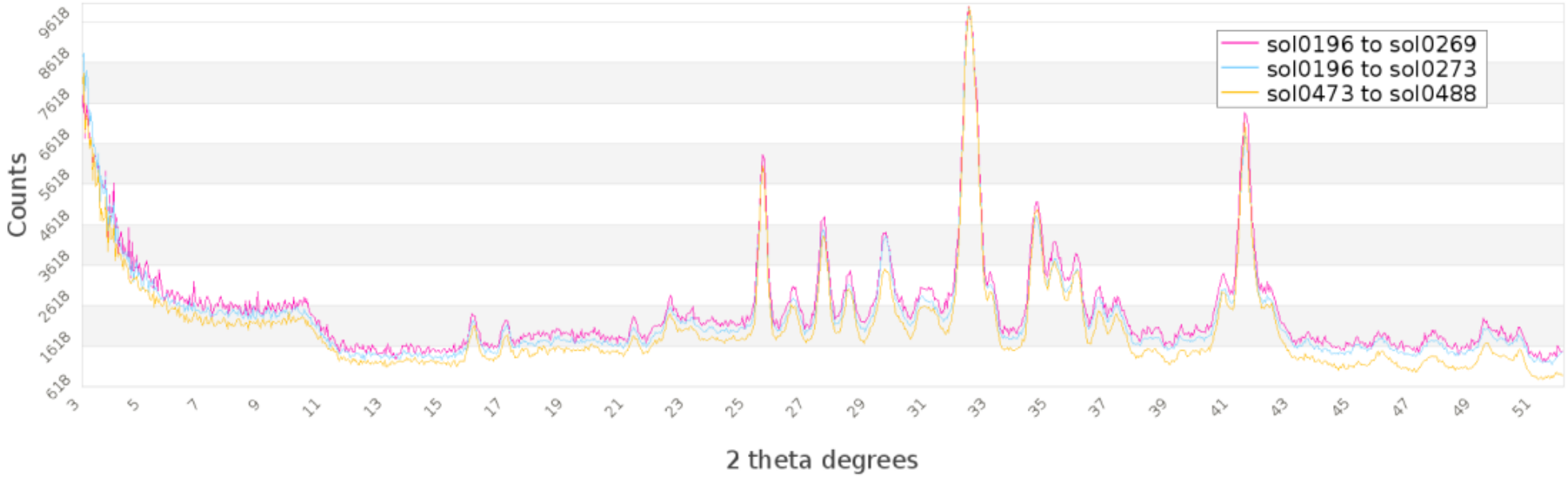
XRD Minor Frames

cma_414856883rda01960060000ch12240p1.lbl.txt

cma_414856883rda01960060000ch12240p1.csv

XRD-RDA Charts

Switch to Dynamic Charts



Mineral Abundances

sol0196 to sol0269

sol0196 to sol0273

sol 0474 to sol 0489

MIN Analysis Name

sol0196 to sol0269

MIN Product_ID

CMA_414856883MIN01960060000CH12240P1

MIN Start_Time

2013-02-23T06:53:49.000

MIN Analysis Description

Example Workflow: Graph Plugin

The graph plugin allows users to plot various data sets graphically in the web browser.

The plugin creates the required fields automatically and allows you to change settings about how the graph will display.

Plugins can also modify individual fields to add specific formatting or display properties to the field.

They can also be used for custom field validation.

As the project is open source, anyone can create a plugin and request inclusion in the next product release.

Render Plugin Settings

Render Plugin: Graph Plugin

The Graph Plugin plots XY data in a line plot on a linear or log chart. A data type with the required fields is required and multiple instances of the XY data files are allowed. If more than one XY file is found, it will be labeled and displayed on a roll-up graph initially. The plugin then allows filtering to a specific file using the meta data provided.

Field Name	Field Description	Field Type	Map to DataField...
Graph File:	A file containing XY values for plotting.	File	Choose a Field...
Pivot Field:	If multiple data sets are plotted, the pivot field will be displayed to allow the user to switch between data sets.	Integer	Choose a Field...

☒ Automatically create DataFields required for this Render Plugin.

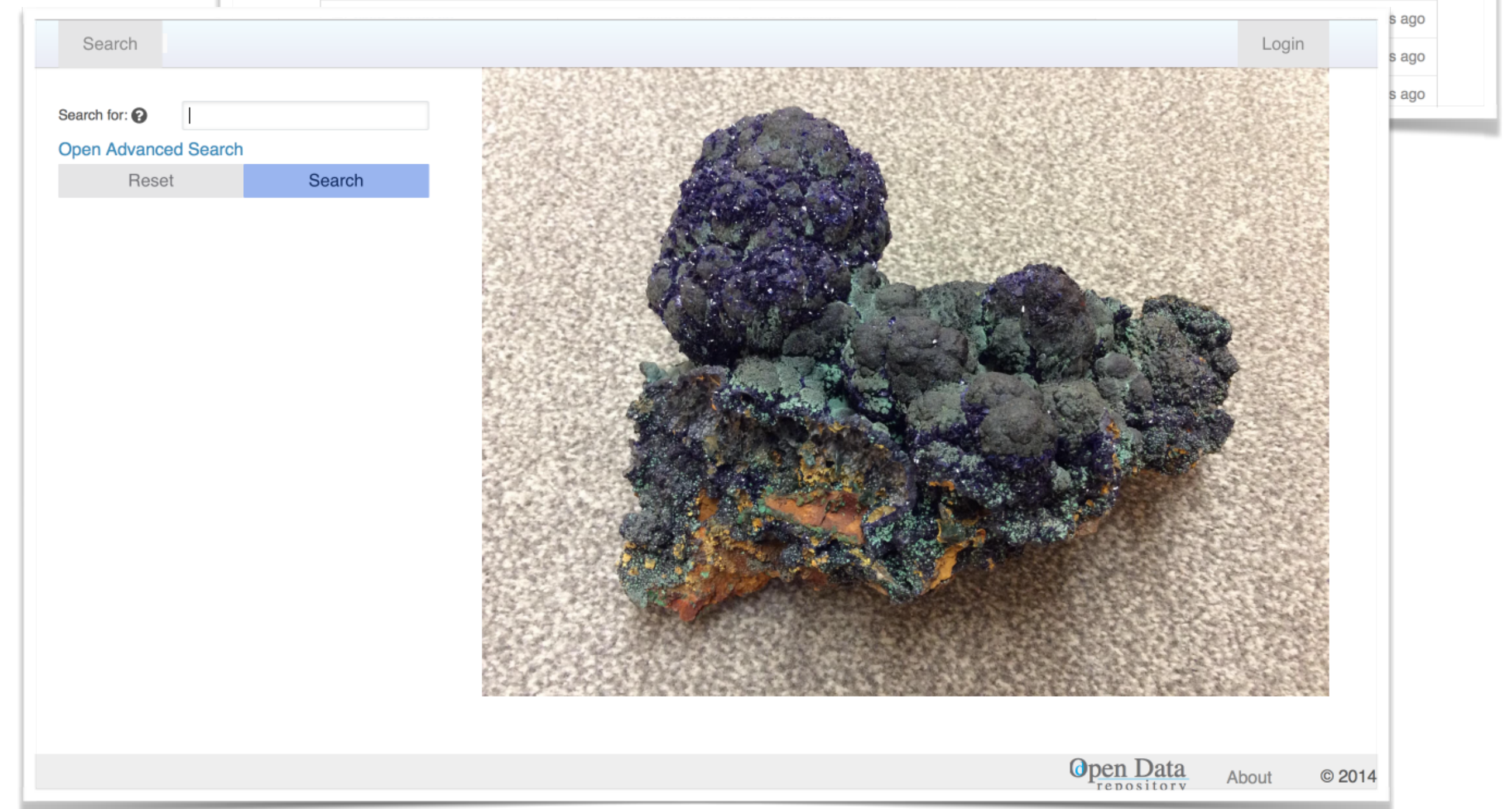
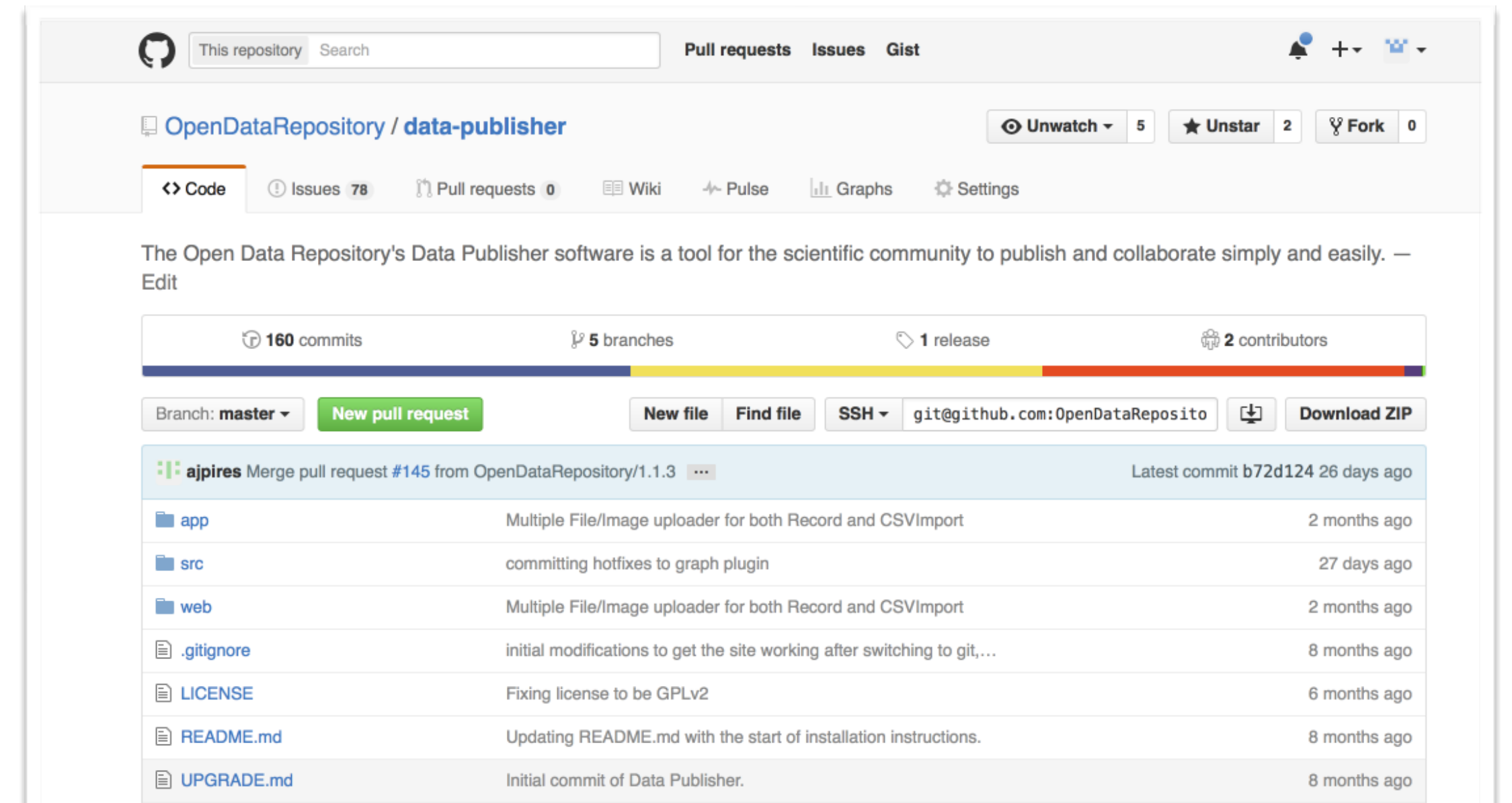
Graph Type ?	XY Scatter	Graph Width ?	900
Graph Height ?	300	X Axis Minimum ?	auto
X Axis Maximum ?	auto	X Axis Direction ?	asc
X Axis Labels ?	yes	X Axis Tick Interval ?	auto
X Axis Caption ?	Width	X Axis Logarithmic ?	no
Y Axis Minimum ?	auto	Y Axis Maximum ?	auto
Y Axis Tick Interval ?	auto	Y Axis Labels ?	yes
Y Axis Caption ?	Height	Y Axis Logarithmic ?	no
Line Width ?	1	Line Colors ?	black
Layout ?	Graphs and Fields are full width.		

Cancel

Save Plugin Settings

Short Term Roadmap

- Source code available on GitHub.
 - <http://github.com/opendatarepository>
- 1.0 beta released, but you'll need our help getting started - we are happy to help.
- Next beta release will occur in early 2016 pending:
 - Creation and testing of Docker instance for ease of administration and deployment.
 - Install script to initialize the database and take users through changing initial passwords, etc
 - Completion of new graphics capability and engine rework.
- Ongoing Pilots:
 - AHED Group at NASA Ames
 - University of Arizona Mineral Museum
 - Collaborative Keck, Sloan, Carnegie, and UofA Mineral Evolution Project
- Looking for pilot participants we can work with to install, configure, and give feedback.



Pilot: Astrobiology Habitable Environments Database

Excellent pilot due to the diversity of fields and data types that astrobiology encompasses.

- Microbiology
- Computer science
- Chemistry
- Geology
- Astronomy
- and more.

Currently, have five databases comprising hundreds of records.

AHED will be defining custom metadata requirements and data structures to facilitate cross-domain searching.

Will be presented Thursday 12/17 at AGU 2015 by Barbara Lafuente in *Data Visualization 1: Data Visualization for Information Exploration and Knowledge Discovery I*

Other Databases

University of Arizona Mineral Museum

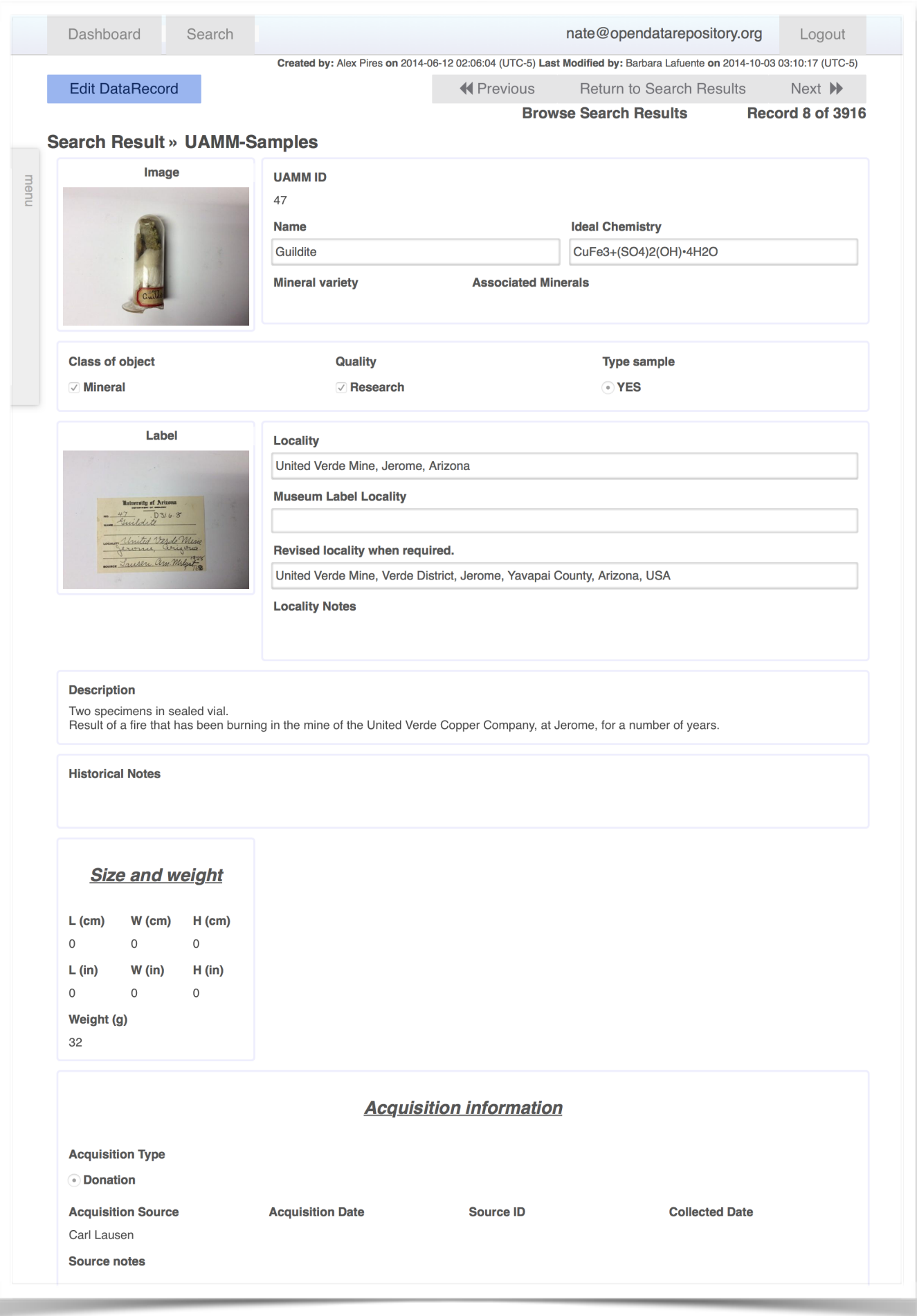
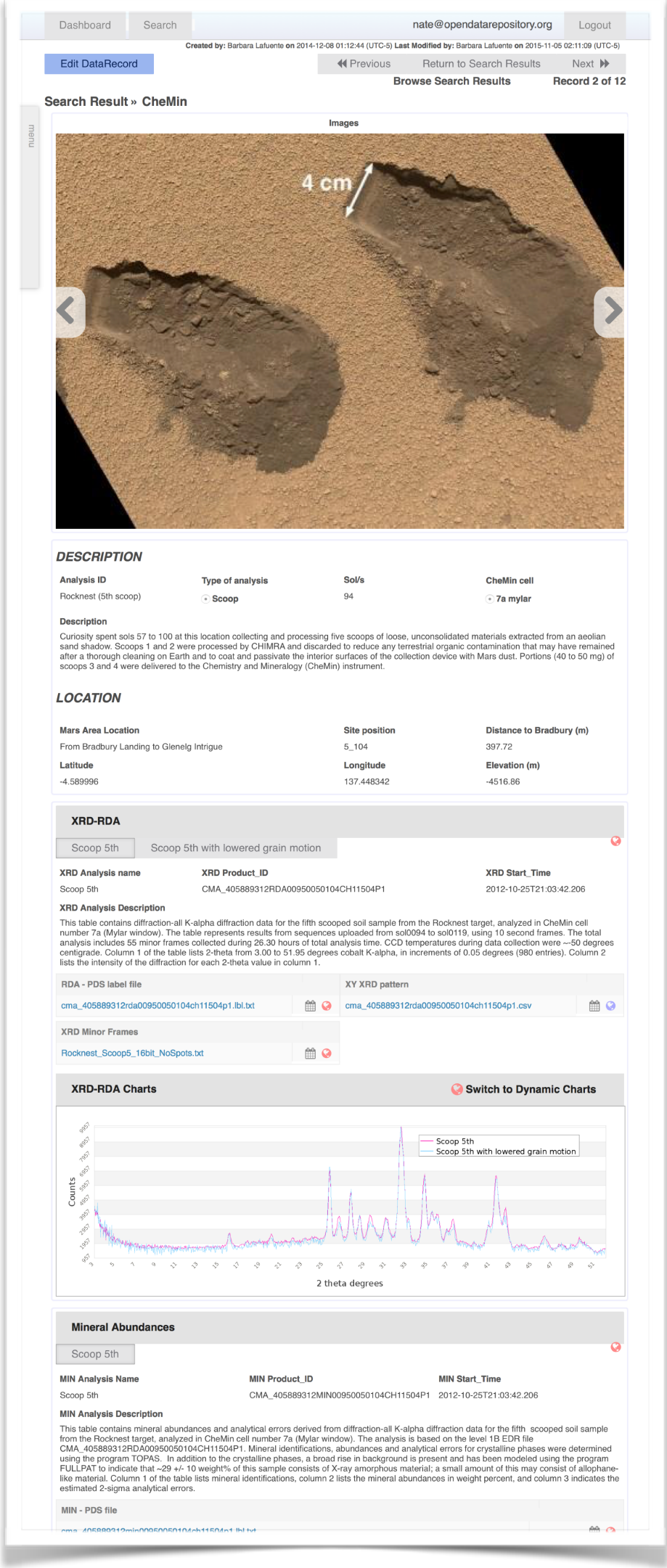
- over 20,000 records
- <http://odr.io/UAMM>

IMA Mineral List (pilot)

- over 5000 minerals
- <http://odi.io/imalist>

Mineral Evolution Database

- Mineral evolution, Robert M. Hazen, et al., *American Mineralogist, Volume 93, pages 1693-1720, 2008*
- 30,000 records and multiple linked databases
- <http://odr.io/MEDCGKB>



Roadmap 2016

- **Q1 2016**

- Docker Installer
- Data Structures using Plugins
- Increased graphing capability - D3
- R/Shiny Integration for live data analysis through web browser.



- **Q2-Q4 2016**

- Data citation system (DOI, PURL, URI, etc.)
- Full REST API for record downloads and edits.
- ODR Hub (ontologies, shared repositories, etc.)
- RDFa implementation
- Ontology-based database creation
- Metadata Standards



Summary

- We have quickly gathered a diverse set of databases that have significant activity.
- Researchers are using them on a daily basis from collection through all phases of research.
- Focus on:
 - Linked data and semantic web integration are fundamental to our long term plans.
 - Citation system is critical before our first full release. Snapshot-based citation generation for data sets or objects.
- As we expand our feature set, expect the system will:
 - provide a very high level of provenance for any data housed in the software,
 - make data management beneficial to the researcher throughout the research process,
 - create living archives that allow citable snapshots of data and summaries of the changes since the snapshot was created.

Thank you.

We gratefully acknowledge the support for this study and the related pilots:

- NASA's Science-Enabling Research Activity (SERA) initiative
- NASA NNX11AP82A, Mars Science Laboratory Investigations
- University of Arizona Geosciences.

Please check out our related talk:

The Astrobiology Habitable Environments Database

B. Lafuente, N. Stone, R. T. Downs, D. Blake, T. Bristow, M. Fonda and A. Pires

Thursday 12/17 at AGU 2015 by Barbara Lafuente

Session: *Data Visualization 1: Data Visualization for Information Exploration and Knowledge Discovery I*

Contact Information

Nate Stone

Open Data Repository

nate.stone@opendatarepository.org

Barbara Lafuente

University of Arizona Geosciences

barbaralafuente@email.arizona.edu

Bob Downs

University of Arizona Geosciences

rdowns@u.arizona.edu

David Blake

NASA Ames

david.blake@nasa.gov

Thomas Bristow

NASA Ames

thomas.f.bristow@nasa.gov