DEMONSTRATING THE OPEN DATA REPOSITORY'S DATA PUBLISHER: THE CHEMIN DATABASE.

N. Stone[1], B. Lafuente[2], T. Bristow[2], A. Pires[3], R. M. Keller[2], R. T. Downs[3], D. Blake[2] , C. E. Dateo[2] and M. Fonda[2]. [1]Open Data Repository, Gray, ME (nate.stone@opendatarepository.org) [2]NASA Ames Research Center, Moffett Field, CA, [3]University of Arizona, Tucson, AZ.

Introduction: Recently, federal agencies have begun mandating that data and results from government funded scientific research be available and useful to the public and the science community. While large, homogenous fields often have repositories and existing data standards (e.g. GenBank), for small communities in multidisciplinary fields publishing and sharing data can be challenging.

In development for nearly four years, the Open Data Repository's (ODR) Data Publisher software has been designed as a collaborative data publication tool for small groups of independent researchers who usually have few options for publishing data that can be utilized within their community.

Objectives: ODR's Data Publisher aims to provide an easy-to-use software tool that will allow researchers to create and publish database templates and related data. The end product will facilitate both human readable interfaces (web-based with embedded images, files, and charts) and machine-readable interfaces utilizing semantic standards.

Characteristics: The Data Publisher software runs on the standard LAMP (Linux, MySQL, Apache, PHP) stack to provide the widest server base available. The software is based on Symfony (www.symfony.com) which provides a robust framework for creating extensible, object-oriented software in PHP. The software interface consists of a template designer where master database templates can be created and customized (Fig. 1). A master database template can be shared by many researchers to provide a common metadata standard that will set a baseline for all derivative databases. Individual researchers can then customize their instance of the master template with specialized fields, file storage, or visualizations that may be unique to their studies. This allows groups to create compatible databases for data discovery and sharing purposes while still providing the flexibility needed to meet the needs of scientists in rapidly evolving areas of research.

The platform facilitates flexible permission sets that enables researchers to share data collaboratively while improving data discovery and maintaining ownership rights. A web-based interface allows researchers to enter data, view data, and conduct analyses using any programming language supported by JupyterHub (http://www.jupyterhub.org). This toolset makes it possible for a researcher to store and manipulate their data in the cloud from any internet capable device. Data can be embargoed in the system until a date selected by the researcher. For instance, open publication can be set to a date that coincides with publication of data analysis in a third party journal.

A CSV import function will automatically generate a template and populate databases from a spreadsheet, allowing users to import large sets of data in a very short time.

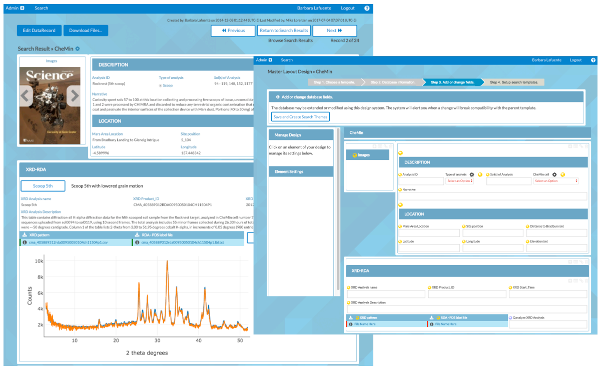


Figure 1. Example of data record view (left) and database template (right) from the CheMin Database.

The CheMin Database: In conjunction with teams at NASA Ames and the University of Arizona, a number of pilot studies are being conducted to assess the needs of individual research groups having disparate projects and data types and to guide the software development so that it allows them to publish and share their data collaboratively. These pilots include the CheMin Database (http://odr.io/CheMin), which contains the data products of the analysis performed by the Chemistry and Mineralogy (CheMin) X-ray diffraction instrument onboard the Mars Science Laboratory, together with tools and procedures for analysis of the data.

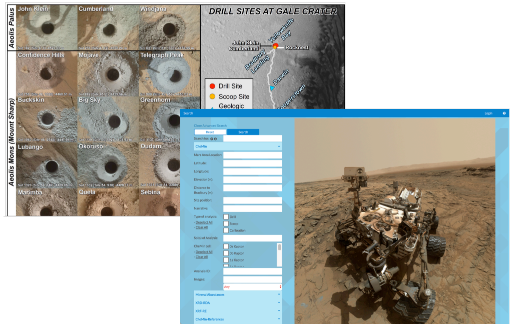
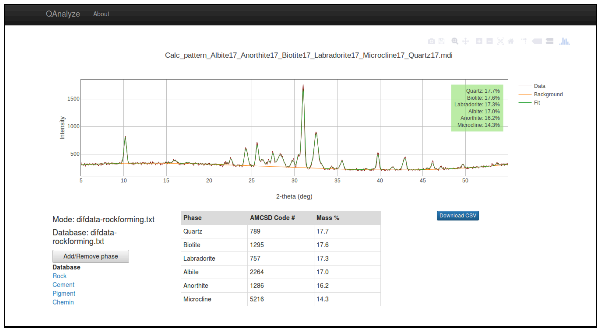


Figure 2. Search interfaces in the CheMin database.

The database benefits from the capabilities of the ODR software which provide a user-friendly interface, where the data are easily accessed using search tools (Fig. 2), visualization using a versatile graphing system, and data downloads in different formats.

The main goal is to provide outside users with the information and data analysis tools that are required to understand and re-analyze the original raw data, replicate experiments, or even perform entirely new studies with different starting hypotheses. Each data record includes: 1) sample description; 2) interactive XRD and XRF patterns with associated metadata and downloadable files; 3) mineral abundances derived from diffraction data; 4) access to the library of CIF files used in diffraction pattern analysis; 5) links to raw data and results from other MSL instruments (such as elemental composition data from APXS) for each of the samples analyzed by CheMin; 6) library of references associated to each analysis; 7) access to the Experiment Data Record (EDR) for each sample; 8) a detailed narrative of how the analysis was performed.

The database also provides access to *QAnalyze* (http://xrd.qanalyze.com/), an automated cloud-based application for quantitative analysis of mineral samples using X-ray diffraction (XRD) (Fig. 3).



Mode: difdata-rockforming.txt

Database: difdata-rockforming.txt

Database
Rock
Cement
Pigment
Chemin

Phase	AMCSD Code #	Mass %
Quartz	789	17.7
Biotite	1295	17.6
Labradorite	757	17.3
Albite	2264	17.0
Anorthite	1286	16.2
Microcline	5216	14.3

Figure 3. Example of analysis in QAnalyze.

Summary: A key feature of databases created using ODR is their ability to change and evolve over time. New data fields can be included and linked without disrupting the basic structure of the database, links can be created for new types of analyses and presentation formats. We are continually adding features and capabilities to the CheMin database (and other databases in the pilot ODR study) as they come available or are seen to be useful.

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