Paper presented at the Physical Samples and Digital Libraries workshop, June 22- 23, 2016, Newark, NJ, USA, held in conjunction with the JCDL 2016 - ACM/IEEE Joint Conference on Digital Libraries

To boldly go: the next generation of the Digital Mineral Library at Curtin University

[Work-in-progress]

Matthias Liffers
Curtin University
Kent Street
Bentley WA, Australia
matthias.liffers@curtin.edu.au

Brent I.A. McInnes
Curtin University
Kent Street
Bentley WA, Australia
directorjdlc@curtin.edu.au

Adam Brown
CSIRO - Mineral Resources
26 Dick Perry Ave
Kensington WA 6151
adam.brown@csiro.au

ABSTRACT

After its success with the Australian National Data Service (ANDS)-funded Major Open Data Collection (MODC) project, Curtin University is working on the next phase of the Digital Mineral Library to integrate data from the SHRIMP ion microprobe, encourage greater data sharing among geoscientists and expand the coverage of the collection.

CCS Concepts

•Information systems → Geographic information systems; •Applied computing → Earth and atmospheric sciences; Digital libraries and archives;

Keywords

Geochemistry, geochronology, open data, open science, geoscience, research data management

1. INTRODUCTION

The Digital Mineral Library[1] at Curtin University is an open-access collection of geochemical datasets that arose from an Australian National Data Service[2] (ANDS) funded Major Open Data Collection (MODC) project. The geochemical datasets are the result of the analysis of 150 mineral samples from the Geological Survey of Western Australia[3] (GSWA) using a TESCAN Integrated Mineral Analyser[4] (TIMA). The resulting datasets have been published in Research Data Australia[5] and the AuScope Discovery Portal[6].

After the successful completion of the MODC project, Curtin University is expanding the Digital Mineral Library to include high value geochronological data produced by the Sensitive High Resolution Ion Microprobe[7] (SHRIMP).

2. MAJOR OPEN DATA

The first phase of the Digital Mineral Library project was a pathfinder to develop the workflows and integrate the systems to permit data publication directly "from the lab workbench". 150 samples of a 2000 sample collection from the GSWA were used to develop this collection. This pathfinder was a collaboration between the Curtin University Library, the John de Laeter Centre[8] (JdLC), Curtin Information Technology Services and external partners GSWA, CSIRO[9], ANDS and AuScope.

2.1 Free and open

Figure 1 is a high-level architecture diagram of the Digital Mineral Library. From the beginning, the project team adhered to principles of openness and freedom. For example, free and open-source software (FOSS) using open standards and APIs were used. Furthermore, some aspects of Linked Open Data were adopted, such as the use of globally-unique identifiers (GUIDs) in the form of HTTP URIs. For each physical sample, the GUID was an International Geosample Number (IGSN) handle. For each dataset, a Digital Object Identifier (DOI) was used.

The transfer of data from the TIMA to the data and metadata storage is handled by the JdLC Laboratory Information Management System (LIMS), which was created for the Digital Mineral Library project. The LIMS serves multiple purposes, including registration of samples and uploading/processing of instrument data.

When a sample is created, its collector/creator can use the LIMS to register the sample with the System for Earth Sample Registration (SESAR), which allocates an IGSN. This IGSN can be embedded into a sample or printed on a container to facilitate easy recognition and tracking of samples.

[1]https://library.curtin.edu.au/about/digital-mineral-library.cfm
[2]http://www.ands.org.au/
[3]http://www.dmp.wa.gov.au/Geological-Survey/Geological-Survey-262.aspx
[4]http://www.tescan.com/en/products/tima/tima-lm
[5]https://researchdata.ands.org.au/
[6]http://portal.auscope.org/portal/gmap.html
[7]http://www.asi-pl.com.au/products/shrimp-ii.aspx
[8]http://jdlc.edu.au/
[9]http://www.csiro.au/

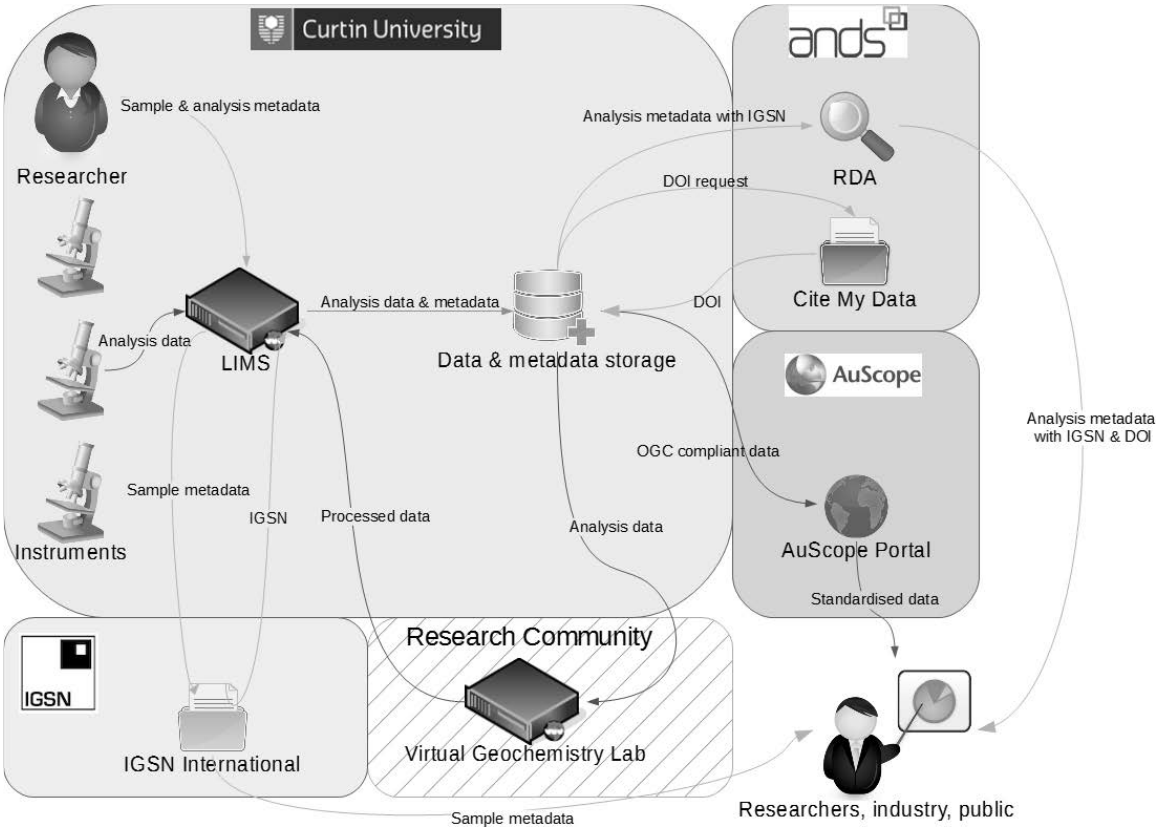

Figure 1: The architecture of the Digital Mineral Library

The LIMS lets an operator in the JdLC register an analysis of a sample and upload the resulting data. A LIMS plugin, particular to each instrument, processes the data and produces files that are suitable for publication. These files are transferred to a data publication server and metadata to a local Geoserver instance and the Curtin Library's ReDBox[10] metadata server.

In the Library, a data librarian processes the metadata record for the dataset to be published. After ensuring that adequate metadata has been provided by the instrument operator, the librarian mints a DOI and publishes the dataset. ReDBox then makes that metadata available to data discovery portals via an Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) interface. OAI-PMH is a common standard used by library systems to syndicate metadata records.

The power of open data could be harnessed in a virtual geochemistry laboratory - a possible cloud-hosted system that permits operators to run virtual experiments using open data. Resulting data could then be re-submitted via the LIMS to be, in turn, openly published.

2.2 From physical sample to digital library

150 samples were received from the GSWA as bags of mineral grains and prepared in the JdLC as "mounts". Each mount contains tens of thousands of microscopic mineral grains, embedded in an epoxy resin disc, complete with a slip of acid-free paper displaying the IGSN and a QR code,

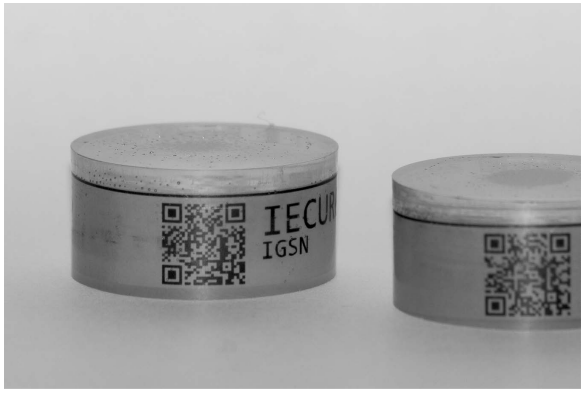

Figure 2: Mineral mounts with embedded QR codes

[10]http://www.redboxresearchdata.com.au/

encoding a link to the SESAR metadata record for that sample. The QR code can be scanned by a mobile device, such as a smartphone, to load the metadata record for that sample.

For sample IECUR000V (**Figure 2**), the embedded QR code leads to, the IGSN metadata record[11] that contains information on the physical sample, including the internal identifier at the Geological Survey of Western Australia (206995H), location, description of the area, age and when the sample was collected.

The matching dataset for IECUR000V is discoverable through multiple data portals, including Research Data Australia, the AuScope Discovery Portal and the Thomson Reuters Data Citation Index.

The metadata record in Research Data Australia [3] provides a clear means of accessing the dataset, as well as information to enable citation and relationships with other datasets in the same collection. Views and downloads for each dataset are tracked. All datasets generated through the Major Open Data Collection were attributed to the JdLC rather than any individual researcher, as they were not generated by a researcher for a specific research project.

The samples and their datasets are available as a layer in the AuScope Discovery Portal[12], providing another avenue for discovery. The AuScope Discovery Portal is a discipline-specific data discovery portal for the geosciences. It enables discovery of physical samples and data in the earth and geosciences.

Finally, all datasets are also available through the Thomson Reuters Data Citation Index, a third avenue of discovery and also a mechanism by which future citations of the datasets can be tracked.

Data use and re-use is enabled using a Creative Commons Attribution 3.0 Australia licence[13] (CC BY 3.0 AU).

2.3 A winning design

The Major Open Data Collection Project team was recognised at the Curtin University Vice-Chancellor's Excellence Awards for Professional Staff 2015, winning the category of Excellence in Innovation[14].

The team attributes this win to the strong collaboration between different sections of Curtin University – the University Library, John de Laeter Centre and Curtin Information Technology Services[1].

3. THE NEXT GENERATION

The second phase of development on the Digital Mineral Library builds on the work undertaken in the first phase. Continuing the partnership with ANDS, AuScope and GSWA, and bringing in new partners Australian Scientific Instruments (ASI; manufacturer of the SHRIMP), Geoscience Australia (GA) and Waymark Systems, the project team seeks to expand capability, encourage a cultural change and extend the coverage. Work in these focus areas is anticipated to be complete by the end of Q2 2017.

3.1 Focus 1: Culture

The Geological Survey of Western Australia is dedicated to making geological information about Western Australia available for research and industry and is a champion for open geological data. The GSWA very quickly agreed to making the first round of Digital Mineral Library data available under a Creative Commons Attribution 3.0 Australia licence.

In phase two, the project's aim is to encourage open data publishing in the JdLC user community. Rather than heavy handedly implementing an open data mandate, the project team is partnering with Waymark Systems[15] to first survey the community members on their current attitudes to open data, publishing original data as well as re-using secondary data. This survey will be an adaptation of the EarthCube Stakeholder Alignment Survey [6], localised for an Australian audience.

The results of the survey will inform the revision of the JdLC SHRIMP terms of use, such as whether there will be any embargo period, the preferred licence and what kind of incentives will exist to encourage open data publishing. The revised terms of use will be taken to the community for further consultation before they are implemented.

3.2 Focus 2: Capability

The first phase of the Digital Mineral Library focussed on the TIMA, which was newly acquired by the JdLC. The second phase of the project will integrate the SHRIMP, an ion microprobe most commonly used in geochronology. The JdLC has two SHRIMPs, which are among the most heavily used instruments in the facility.

Integrating the SHRIMP into the Digital Mineral Library workflow will involve creating a processor plugin for the LIMS to accept the SHRIMP data and create outputs suitable for data publication. For this, the project team is partnering with ASI, the manufacturer of the SHRIMP.

The project team is committed to properly acknowledging dataset creators using best-practice. The ORCID standard [2] has quickly achieved dominance in the researcher identifier space, partially due to its open and publisher-agnostic nature. ORCID iDs will be used as GUIDs for dataset creators, closing the loop on the Linked Open Data principles adopted at the beginning of the project.

The final capability to be implemented is formal provenance linking using a standard such as PROV-O [4]. The Digital Mineral Library already links datasets to their originating samples, but not capture any descriptions about the analyses themselves, such as particular protocols or standards used. A formal provenance system will record this metadata. The use of IGSNs and DOIs in the first phase of the Digital Mineral Library will simplify this integration.

3.3 Focus 3: Coverage

The 150 samples used to first populate the Digital Mineral Library were collected by the GSWA over several decades. Since the GSWA is by definition focussed on the geology of Western Australia, the resulting mineral map represents the geochemistry of only Western Australia (**Figure 3**).

Geoscience Australia will be contributing data to the Digital Mineral Library after analysis of samples gathered originally from the North West Shelf [5]. These samples will be initially analysed with the TIMA, with further analysis undertaken with a laser microprobe.

4. CONCLUSIONS

[11] https://app.geosamples.org/sample/igsn/IECUR000V
[12] http://portal.auscope.org/portal/gmap.html
[13] https://creativecommons.org/licenses/by/3.0/au/deed.en
[14] https://youtu.be/vRfDTORRZIc
[15] http://waymarksystems.org/

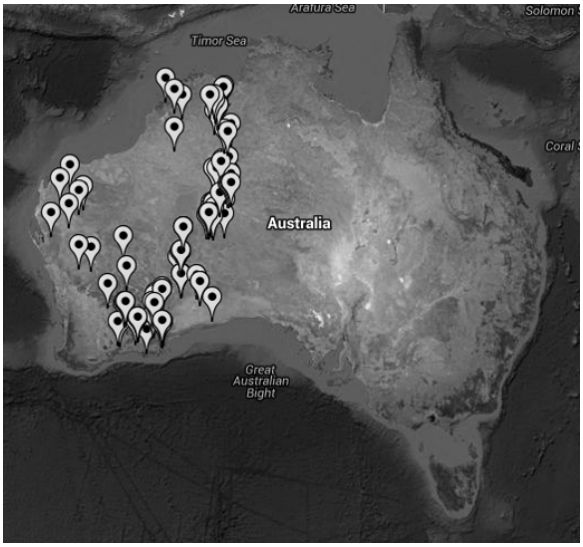

Figure 3: A clear bias in first phase coverage

The Digital Mineral Library achieved successful through strong collaboration and dedication to openness: open science, open access, open source and open standards. For the next phase of the project, the Digital Mineral Library team has forged new partnerships to focus on changing culture, improving capability and expanding coverage. This phase of the project is expected to be complete by July 2017.

5. ACKNOWLEDGEMENTS

This e-Research project is supported by the Australian National Data Service (ANDS) and AuScope through the National Collaborative Research Infrastructure Strategy Program. The hardware component of the project was funded via the Australian Research Council with support from Curtin University, the Geological Survey of Western Australia, the University of Western Australia and Murdoch University.

This project is jointly run between the Curtin University Library and the John de Laeter Centre.

6. REFERENCES

[1] Australian National Data Service. The aware winning team. *share*, 24:6–7, 2016. Retrieved 10 May 2016 from http://www.ands.org.au/news-and-events/share-newsletter/share-24/the-award-winning-team.

[2] L. L. Haak, M. Fenner, L. Paglione, E. Pentz, and H. Ratner. Orcid: a system to uniquely identify researchers. *Learned Publishing*, 25(4):259–264, 2012. Retrieved 8 May 2016 from http://doi.org/10.1087/20120404.

[3] John de Laeter Centre. Tima energy dispersive system analysis of a sample of igneous rock igsn:iecur000v (raw sem). http://doi.org/10.4225/06/5537277BB14A9, 2015. Retrieved 29 April 2016.

[4] T. Lebo, S. Sahoo, and D. McGuinness. Prov-o: The prov ontology. https://www.w3.org/TR/prov-o/, 2013. Retrieved: 2016.04.29.

[5] C. J. Lewis and K. N. Sircombe. Use of u-pb geochronology to delineate provenance of north west shelf sediments, australia. In M. Keep and S. Moss, editors, *The Sedimentary Basins of Western Australia IV: Proceedings of the Petroleum Exploration Society of Australia Symposium*, 2013.

[6] S. Ramdeen. Developing cyberinfrastructure for earth science: an opportunity for collaboration. *D-Lib Magazine*, 19(5/6), 2013. Retrieved 8 May 2016 from http://doi.org/10.1045/may2013-ramdeen.