Data and Services Discovery projects - Transformative Data Collections

Final report

Title

Bringing Long-Tail Microscopy and Characterisation Data into the Light

Approach

The organisations involved in this Project are:

- Microscopy Australia (MicroAU)
  - Nodes at the Centre for Microscopy and Microanalysis (CMM) at the University of Queensland (UQ); Australian Centre for Microscopy & Microanalysis (ACMM) at the University of Sydney (USyd); and the Centre for Microscopy, Characterisation and Analysis (CMCA) at the University of Western Australia (UWA)
- National Imaging Facility (NIF)
  - Nodes at the Centre for Advanced Imaging (CAI) at UQ; and CMCA at UWA
- Australian Research Data Commons (ARDC)
- UWA Library

The Project Team comprised:

1. Data & Informatics Committee representatives from the participating MicroAU nodes: Roger Wepf (UQ, Head of the Committee), Ryan Sullivan (USyd), Matt Foley (USyd) and Andrew Mehnert (UWA);
2. NIF Informatics Fellows from the participating NIF nodes: Aswin Narayanan (UQ) and Andrew Mehnert (UWA);
3. Lisa Yen (Chief Operating Officer of Microscopy Australia);
4. Abby Asomani (UWA Library Information Specialist - Research Publication & Data Services);
5. Mingfang Wu (ARDC);
6. Alexander Joos (Data Management, Analysis & Visualisation, CMCA, UWA)

Project Team and Writing Group meetings

The Project formally commenced in August 2019. A weekly Project Team Zoom meeting series was initiated on 6 August. A total of 12 such meetings were held, during which deliverables were discussed, action items assigned/reported, and project planning undertaken. Project documents, including deliverables, were shared in a Google Drive folder. Team members collaborated on the documents using Google Docs. A writing group involving Abby Asomani (part-time secondment from the UWA library), Aswin Narayanan (sub-contracted to this project) and Andrew Mehnert (Project Manager) engaged in additional face-to-face and Zoom meetings as needed to progress the deliverables.

Node visits

Deliverable 3 involved cataloguing instrument-generated file types and available metadata extraction and file conversion tools and services across MicroAU and NIF. The exercise included not only input from the Project Team but also from instrument managers within the ACMM, CMM and CMCA. To ensure accurate and complete collection of this data, Aswin Narayanan visited each site (including 1 week at CMCA and 1 week at ACMM), consulted with each instrument manager and performed all data entry.

Progress Report and Summit Presentation

Delays associated with contracting between UWA and ARDC, and subcontracting between UWA and UQ, delayed the start of this Project. For this reason the ARDC granted permission for the project to continue through to the end of December 2019. A Progress Report was submitted to the ARDC on 8 October and a presentation given at the ARDC Data and Services Summit at the Brisbane Convention Centre on 21 October 2019.

Project Outputs

The Table below lists the seven tasks associated with the proposed approach outlined in the original application, and the associated outputs. The outputs are archived in the National Imaging Facility's CRM and are available upon request. Live versions of the documents can also be accessed via the hyperlinks in the Table.

Activity	Output(s)
1. Investigate/propose a sufficiently flexible data model	Report and Data Model Table
2. Catalogue existing metadata standards and vocabularies for characterisation instruments and identify where there are gaps	Report

3. Catalogue file types across all MA and NIF instruments and also available metadata extraction and file conversion tools and services and identify where there are gaps.	Survey and summary
4. Investigate/propose a data packaging specification suitable for interoperability	Report
5. Investigate/propose a standardised protocol for collecting quality data from characterisation instruments	Report
6. Catalogue/evaluate suitable cloud-based platforms	Survey and summary
7. Investigate/propose suitable, community agreed, licenses for data publishing	Report

FAIR

- Activities 1-3 address rich metadata for data and associated entities (e.g. instruments). Unique persistent identifiers (e.g. DOIs, handles, RAiDs) can be associated with these and records deposited in Research Data Australia to make them findable.
- Activities 1, 5 and 6 relate to the deposit of quality instrument data into a trusted data repository to make it accessible.
- Activity 4 addresses data interoperability.
- Activity 7, together with the researcher-controlled data publishing model of the MyTardis platform (http://www.mytardis.or g) , addresses reusability.
- Please refer to the attached FAIR assessment spreadsheet.

Collaboration and coverage

This project is a collaboration between two national characterisation capabilities: Microscopy Australia (MicroAU) and the National Imaging Facility (NIF). The two collaborating NIF nodes

(UQ and UWA) have established repository services (developed as part of the 2017/2018 ANDS-RDS-NIF Trusted Data Repositories project). At the end of this project, the resulting documentation and shared experience will enable MicroAU nodes to deploy their own services (as part of a federation of repository services). Moreover, the shared outcomes will help both MicroAU and NIF to better address the “I” and “R” of FAIR.

Sustainability

As noted above the resulting documentation and findings are available on request from NIF, and can also be accessed via the hyperlinks in the Table above. They represent a resource not only for NIF and MicroAU, but also the wider community,

The existing community of NIF Informatics Fellows, the MicroAU Data & Informatics Committee, and the new appointment of Informatics Fellows across several MicroAU nodes provides an opportunity to evolve these documents and standards, as well as to implement sustainable federated repository services for characterisation instrument data. Indeed, several of these ideas form the core of the recently awarded ARDC Platform s proposal “The Australian Characterisation Commons at Scale” involving Monash (lead), MASSIVE, USyd, UQ, UWA, UNSW, UoW, UMelbourne, AARNET, MicroAU and NIF.

Learnings

- That much of the most crucial information required for effective long-term curation and reuse must be captured at the conceptualisation and collection stages[1]. In the case of characterisation instrument data this means that we need to capture this metadata at the time of upload to the repository service. This includes both the implicit metadata and explicit metadata. The former is generated by the instrument alongside the data itself; e.g. embedded in the resulting data files (e.g. TIFF) or within log files. The latter, on the other hand, requires human input. For example, when uploading a group of related files (dataset) from an instrument to a repository, the user can provide metadata about the dataset itself; e.g. information about the specimen or the experiment. NIF and MicroAU can provide national leadership here by defining templates/schemas for such metadata for individual instruments or groups of similar instruments; e.g. as key-value pairs.
- That the variety of instrumentation and domain-specific needs (e.g. clinical vs preclinical imaging) across the characterisation community means that MicroAU and NIF will need to support a variety of data repository platforms (e.g., OMERO, XNAT, MyTardis).
  - NIF and MicroAU can, however, agree upon a common data packaging standard to facilitate interoperability between repositories and compute services such as the CVL.
  - Rather than trying to write metadata extractors (for implicit metadata) for every platform, it would make sense to invest in developing a cloud-based service that the platforms could query.

- That services are needed for persistent identifiers such as RAiD, DOI, ORCiD; metadata standards and vocabularies; and metadata extraction and data transformation. The ARDC should support the development and delivery of these services.

Impact

We believe the resulting documentation and findings—data model, data packaging specification, standardised protocol for collecting quality data, licenses for data publishing, catalogues, evaluations and gap analyses—will establish the groundwork for developing a national network of federated shared data repositories, based on the FAIR data principles, for MicroAU and NIF (interoperable) for data acquired from >200 instruments. It would support new research and industry opportunities for multi-instrument / multi-technique characterisation across both NIF and MicroAU, as well as ANSTO (many of our users acquire Australian Synchrotron data). It would also support multi-site projects and the possibility for creating new, scientifically valuable collections; e.g. a national digital repository of CT-scanned biological specimens from museums, or a national medical image skin cancer database.

The lack of open standards in the characterisation community, exemplified by electron microscopy, leads to hundreds to thousands of man hours being wasted in finding and sharing data, converting data between formats, seeking missing parameters and fixing missing values. This is an absurd situation in the digital age. This situation needs to be resolved if characterisation is to broadly transition from a qualitative to a quantitative scientific discipline. We believe the Project outcomes will be a catalyst for this transition.

Report prepared by: Andrew Mehnert, CMCA, UWA
Date: 20/12/2019

[1] http://www.dcc.ac.uk/node/9554