Data and Services Discovery projects - Transformative Data Collections

Title

Understanding and Creating Instrument Generated Data Collections

Approach

Activities Undertaken:

- Targeted in-person workshops with the following stakeholders:
  - Nebojsa Tomasevic, Simulator Technician, Monash Accident Research Centre
  - Jay van Schyndel, Research DevOps Engineer, Monash eResearch Centre
  - Asadul Haque, Senior Lecturer in Geomechanics Engineering, Department of Civil Engineering
  - Michael de Veer, Head of Pre-Clinical Imaging, Monash Biomedical Imaging
  - Stephen Firth, Manager, Monash Micro Imaging
  - Alex Fulcher, Senior Microscopist, Monash Micro Imaging
  - Robert Goode, Proteomics Research Assistant, Monash Proteomics & Metabolomics Facility
  - Neil Dickson, Manager, Research Infrastructure, Monash Library
  - David Abramson, Director, Research Computing Centre, University of Queensland
  - Graham Galloway, Chief Executive Officer, National Imaging Facility
  - Roger Wepf, Director of the Center for Microscopy and Microanalysis, University of Queensland
  - Aswin Narayanan, Research Support Software Developer, University of Queensland
  - Luke Visser, Emerging Technology Manager, Agilent Technologies

Combined, these facilities have integrates 79 instruments with Store.Monash[1], storing more than 886 TB of data and serving more than 1,000 distinct researchers.

- Data collection using Store.Monash as the main use-case.

FAIR

Please check the attach FAIR assessment spreadsheet.

Collaboration and coverage

This project is a collaboration between:

[1] Monash instance of MyTardis

- **Monash University**: Amr Hassan, Wojtek Goscinski, James Wettenhall
- **University of Queensland**: David Abramson, Aswin Narayanan
- **University of West Australia**: Andrew Mehnert

The coverage of this report extends to include discussions with Graham Galloway National Imaging Facility (NIF).

Sustainability

The outcome of this project will contribute to the implementation roadmap of MyTardis.This roadmap will be discussed with the community during the annual MyTardis workshop at eResearch Australasia 2019 (Friday 25th October 2019 - Brisbane).

MyTardis is currently a self-sustaining project: Monash University sustains the current code base, and UQ, and UWA provide modest code contributions and expertise.

Learnings

Our stakeholder consultations and workshops focused on “How Instrument data management platforms can facilitate and support making data more FAIR?”. The following points summarize the learnings and the information we gathered during these consultations:

How to make the data findable and accessible?

Capturing the data at its source is the best approach to ensure that the data is protected and managed appropriately. It reduces the possibilities of data loss or corruption and ensures that the appropriate data retention is applied. However, the instrument data management platforms should not assume that all the data captured at this early stage is useful and should be retained. Some facilities and researchers estimated that around 80%[2] of the data collected from their instruments might not contribute directly to a final scientific outcome, but is still essential to the overall scientific process and must be captured and managed. While this might not be the case for all the instruments, Instrument data management platforms should consider this as a possibility and provide the tools to support handling the case where permanent data storage is not required. This includes:

- lowering the entry barrier at the experiment and data collection stage and automate to a great extent the process of data gathering. Adding complicated meta-data gathering steps at the first stage will impede and slow down the data gathering activity. At this stage of data gathering, the focus should be on the automated data extraction mechanisms such as extracting meta-data from the file headers, from the instrument used (some instrument manufactured at moving to adopt the Allotrope data format[3]), or from the lab/instrument booking systems (e.g Agilent iLab at Monash[4]);

[2] At this stage, this estimate is anecdotal with no quantitative data to support it.
[3] https://www.allotrope.org/solution
[4] https://www.agilent.com/en/products/lab-management-software

- providing an easy-to-use mechanism for the researcher to indicate if the data collected will be useful or not. Based on this an appropriate data retention policy should apply; and
- providing an easy-to-use mechanism to indicate the relationship between different data files or data sets to highlight basic data processing operations such as file format conversion.

These features will ensure the quality of the data collections and ensure that the productivity of the researchers is not negatively impacted by the data management processes.

To further improve the data management process, the role of instrument data management platforms should be extended to manage the full data life-cycle. This includes managing derivative datasets and maintaining the link between these datasets and the original data collected to provide the appropriate data provenance. During this data transformation process, the data management platforms should provide the ability to integrate with other platforms such as Electronic Lab Notebook (ELN), associated method paper(s), upload documents to describe the process, group files and datasets into projects, add ethics approval details, and associate tags to files and datasets as a mechanism of grouping.

Up to this stage, the data is only accessible to the data owners and their collaborators. The instrument data management platforms should provide search functionality that facilitates making this data finable while maintaining the data accessing rights.

The process of data publishing should be fully governed by the researchers. Within all the consultations conducted, there has been agreement that the raw datasets and the derivative datasets might not be in a format that is ready for publication. The data owners want to maintain full control of the data publishing process and how the data will be presented to their research community. This includes the ability to rename files/datasets, include/exclude specific files, include/exclude meta-data fields, select a specific license for re-using the data, and specify an appropriate credit or acknowledgement text.

The role of the instrument data management platforms in this process should be to facilitate and encourage the publishing process including providing a unique, citable and persistent identifier for the published dataset, and providing customizable pre-defined meta-data schema. The type of meta-data associated and the completeness of this metadata should be determined and assessed either by the journal or conference where the scientific results will be published, the funding agency or the scientific community.

At this stage, the instrument data management platforms can facilitate making the data searchable by the public via enabling data indexing by data or internet search engines (e.g. Google Public data or Google Scholar). To facilitate such indexing, instrument data management platforms should support the ability to dynamically group public datasets or data files into collections based on the circumstance of data capture, and the instrument used, the subject studied is the key to improve the data reusability. This facilitates exploring such datasets and make it more search engines friendly.

How to make data interoperable?

The instrument data management platforms have limited control over the data format exported from the instrument. However, instrument data management platforms should provide the ability for file format conversion (or export) to facilitate data re-use and interoperability. These conversion tools should be supported as plugins or add-ons managed and maintained by the different research communities. The role that the instrument data management platforms play in this case is to gather all of these tools in the same workspace and support on-the-fly data conversion to lower the entry point.
At the platform level, instrument data management platforms should provide secure and easy-to-use standard APIs to facilitate the integration between different workflow tools or custom code and these platforms. The availability of such integration capabilities will encourage and facilitate integrating the data management platforms in the researchers' data analysis and processing as a data source and destination. Additionally, this will facilitate the interoperability between different data management platforms and the ability to export data packages from one platform to another.

How to facilitate data reusability?

In general, most of the instrument data is captured under very specific conditions, looking at a particular subject, and part of a specific experiment. Reusability of such datasets is directly proportional to the comprehensiveness and the completeness of the meta-data and provenance information associated. The metadata model used should support incorporating information about the published paper(s) and links to code repositories to maintain the association between the data and the code used to process and analyze it.

How to increase the trustworthiness of the data?

Most of the instruments maintain a set of quality control datasets. At the moment, these datasets are maintained outside the instrument data capture framework. This data might include reference images, graphs, Signal-to-noise ratio variation over time or just pass/fail sheets. The ability to incorporate these datasets and cross-reference them with the data gathered will facilitate the process of data validation and increase the trustworthiness of the data collected. Research data captured should easily reference the appropriate quality control data.

How to improve data management and retention?

The policies governing data management and retention varies between different organizations and in some cases between different instruments within the same organization. Within Monash, the instrument management team is responsible for data management and retention on behalf of the researchers. Hence, storage allocation and collection management are managed by this team. Within UQ, data management and retention are managed by the researchers themselves. They are responsible for storage allocation and collection management. Each of these approaches has its advantages and disadvantages. The instrument data management platforms should aim to support a flexible implementation that satisfies the needs of both approaches. We think there is an opportunity to implement a mixture between the two approaches, where the

data is a shared responsibility between the instrument management team and the researchers for a defined time frame to ensure that the data is captured at the source. After completion of the experiment, data management and governance becomes the responsibilities of the researcher. At the transition stage, the data management platform can enforce a minimum set of meta-data such as details about the grant/ ARC proposal/ Project proposal, the required retention period, and who can have access to this data. Having these additional details will enable the data management platform to implement predefined data retention and archiving policies.
The additional information provided by the researcher at the transition stage, will enable accurate reports at the institution level and allow the research to indicate if these datasets are useful or no longer needed. Additionally, it enabled the platforms to implement predefined sharing agreement with other users or organizations if applicable.

What does the ARDC need to address at the national level to make the process easier?

The culture of sharing and publishing data is still the main barrier, especially in some domains such as life sciences. While many improvements were suggested in this report to improve the instrument data management platforms, we don't think the barrier for data sharing and reusability is a technical barrier. Incentive exists for data sharing at the publication stage but not prior to that. Activities aiming at changing such culture and encouraging data sharing should be the focus.
On the other hand, the features and functionalities mentioned above require sustained investment at the national level to improve and integrate platform capable of providing these functionalities to the national and international community. No single institution is capable of undertaking the improvements identified. However, there was broad consensus that a national approach can lead to significant improvements.

Impact

This work will be a major factor within the MyTardis implementation roadmap for the next two years. MyTardis development is a national effort to improve instrument data capture and management. Across the three participating universities, MyTardis manages 1PB+ of data, integrates over 100 instruments, and serve 1000+ researchers. Recently, Monash University has been working with the instrument manufacturer Agilent to enhance the functionalities of MyTardis and the integration with different third-party tools. This report will be discussed within the annual MyTardis workshop at eResearch Australasia 2019.
We believe this report will be valuable to publish publicly as a set of recommendations to data management initiatives.

Report prepared by: Amr Hassan and Wojtek Goscinski, Monash University
Date: 14/10/2019