ORCHESTRATION AND MANAGEMENT OF DATA GENERATED BY BIG-DATA ELECTRON MICROSCOPY INSTRUMENTS: A DISCOVERY REPORT

BIG-DATA ELECTRON AND CORRELATIVE MICROSCOPY FROM INSTRUMENT TO PUBLICATION

David POGER

Jay VAN SCHYNDEL

Hoang NGUYEN

Joshua SILVER

Wojtek J. GOSCINSKI

AUSTRALIAN CHARACTERISATION COMMONS AT SCALE

MAY 2021

Dr David POGER
Microscopy Australia

Jay VAN SCHYNDEL
Monash University

Dr Hoang NGUYEN
The University of Queensland

Joshua SILVER
University of Wollongong

Prof Wojtek J. GOSCINSKI
Monash University

Contacts:

Dr David POGER
david.poger@micro.org.au

Prof Wojtek J. GOSCINSKI
wojtek.goscinski@monash.edu

This report and the recommendations made herein were presented to the stakeholders of Work Package 4 of the Australian Characterisation Commons at Scale project (ACCS) in October 2020 and discussed in order to determine deliverables for Work Package 4. Then, based on the stakeholders' feedback, the proposed deliverables were presented to the Scientific Advisory Groups that oversees Work Package 4 and to the ACCS Steering Committee to seek their endorsement.

ARDC

Australian Research Data Commons

This project is supported by the Australian Research Data Commons (ARDC) and the following partners. The ARDC is enabled by NCRIS.

MONASH University

MICROSCOPY AUSTRALIA

National Imaging Facility

aarnet

MASSIVE

QCIF

PAWSEY supercomputing centre

THE UNIVERSITY OF SYDNEY

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

UNSW SYDNEY

THE UNIVERSITY OF WESTERN AUSTRALIA

UNIVERSITY OF WOLLONGONG AUSTRALIA

THE UNIVERSITY OF MELBOURNE

Flinders UNIVERSITY

University of South Australia

RMIT UNIVERSITY

Contents

Acknowledgements

THE UNIVERSITY OF
MELBOURNE

Melbourne Advanced Microscopy Facility

A/Prof Eric Hanssen

THE UNIVERSITY
OF QUEENSLAND
AUSTRALIA

Centre for Microscopy and Microanalysis

Prof Roger Wepf
Dr Rubbiya Ali
Dr Hoang Nguyen

THE UNIVERSITY OF
SYDNEY

Sydney Microscopy and Microanalysis

Dr Matthew Foley

THE UNIVERSITY OF
WESTERN
AUSTRALIA

Centre for Microscopy, Characterisation and Analysis

A/Prof Martin Saunders
Dr Andrew Mehnert
Dean Taylor

NEW YORK STRUCTURAL BIOLOGY CENTER (USA)

Simons Electron Microscopy Center

Dr Bridget Carragher
Clint Potter

MONASH
University

Monash Ramaciotti Centre for Cryo-Electron Microscopy

Dr Georg Ramm
Hari Venugopal
Jay van Schyndel

Monash Centre for Electron Microscopy

A/Prof Matthew Weyland
Dr Peter Miller

UNSW
SYDNEY

Electron Microscope Unit

Prof Richard Tilley
Dr Nicholas Ariotti
Luc Betbeder-Matibet
Kiho Cho
Frank Li
Jake Surman
Dongming Zheng

UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Cryogenic Electron Microscopy facility

Dr Simon Brown

UNIVERSITY OF
ILLINOIS
URBANA-CHAMPAIGN

Materials Research Laboratory

Dr Mauro Sardela
Dr Timothy Spila

Abbreviations

APT	atom-probe tomography
CVL	characterisation virtual laboratory
cryo-EM	cryogenic electron microscopy
EM	electron microscopy
FIB	focused ion beam
HPC	high-performance computer (or compute or computing)
LSFM	light-sheet fluorescence microscopy
MRI	magnetic resonance imaging
SEM	scanning electron microscopy
STEM	scanning transmission electron microscopy
TEM	transmission electron microscopy
XRM	X-ray microscopy

Executive summary

For this report, representatives from academic facilities in Australia and overseas that operated, or were planning to operate, electron microscopy and correlative light–electron microscopy instruments that produced large volumes of data, were interviewed. The interviews aimed to collect information on how the facilities had set up their data workflows for data capture, data transfer, data movement, data storage and overall data orchestration, which tools they used for data processing (including the supporting infrastructure), and how data and metadata were managed. Based on this survey, the report provides a review of the informatics and data-management landscape at Australian facilities, including tools, methods and procedures currently in use or commented on during interviews. Furthermore, the report proposes recommendations to the stakeholders of the Australian Characterisation Commons at Scale (ACCS) project and its Work Package 4 for tools, methods and procedures that have been considered interesting, promising or relevant to examine further over the first year of Work Package 4 (2021), and guide ACCS work packages into the future.

The interviews were synthesised and assessed around four key areas for which common features, tools, procedures, challenges and gaps were identified: data movement, data processing, data management (including metadata) and data orchestration. Note, this report focuses on the IT infrastructure challenge of big-data-producing instruments, in particular in the case of electron microscopy, and makes recommendations accordingly. Aspects such as researchers' training or detailed modality-specific techniques and algorithms are important but are not directly addressed under this specific discovery activity.

Data movement involves a combination of hardware and software involved in the data lifecycle from the creation or generation of data at an instrument to data capture, storage, processing and archival. The underlying network infrastructure and the tools used to move data along a workflow are the two essential components of data movement to consider when improving and optimising the transfer of large data volumes across or between facilities and institutions. The interviews showed that data movement consisted of a range of routine tasks that required human intervention for actuation or *ad-hoc* adjustments. Ultimately, most of data movement should be based on optimised and automated workflows with limited human intervention.

Data processing is a series of steps to extract meaningful and significant information from raw data. Some processing steps can happen on-the-fly as data are generated from the instruments, while other steps are performed offline. A range of data-processing software packages were identified during the interviews. In order to deal with the volume and velocity of data, facilities indicated that data processing was often carried out on high-end workstations or high-performance computers. The report covers two aspects of data processing: software packages and the supporting computing infrastructure.

Data management ensures that the quality and the integrity of research data are maintained over the data lifecycle (collection, organisation, storage, preservation and sharing) and that legal, ethical, governance and funding requirements are met. Furthermore, good practices in data management can maximise the value of research data. In general, all the facilities applied some level of data management (in particular for data storage) but general guidelines, especially on data retention and deletion, were needed considering the increasing volumes of data generated by instruments. The report also deals with tools for data management as well as standards and guidelines for the capture of metadata, the use of persistent identifiers and the promotion and adoption of the FAIR data principles across facilities.

Data orchestration focuses on the overall process from data capture at the instrument, movement of data to storage for processing and then data management for longer-term storage, including provisions for discoverability and accessibility. In this report, data orchestration is defined in the context of big-data microscopy, data states are discussed, tools potentially suitable to automate the orchestration processes are suggested, and two example workflows are presented.

Below is the list of the key findings and recommendations in the four areas covered by the report: data movement, data processing, data management and data orchestration.

SUMMARY OF THE KEY FINDINGS AND RECOMMENDATIONS

Data movement

Key findings

- All facilities expressed significant data-movement requirements and challenges.
- Fast and efficient data transport was critical to the operation of the experiment or instrument overall.
- A range of tools and protocols were described but no single tool was identified as a solution because of wide-ranging requirements and infrastructure.
- Clear end-to-end understanding of the network was important.
- Few higher-level data transport services were applied.

Recommendations

1. Develop reproducible baselining or benchmarking of file-transfer performance, and a process for re-baselining when components of the network infrastructure change which would affect the ability of data-movement tools to continue to work efficiently. Share and distribute the information across the ACCS network.
2. Investigate and prototype a higher-level transport service for data transport. Usability and user experience will prove important, as will long-term support.
3. Develop optimised and automated workflows for data movement where relevant in order to minimise human intervention for routine tasks.
4. Establish reproducible processes for data archiving to ensure that computing facilities are able to continue processing new data.

Data processing

Key findings

- Data-processing requirements were significant and growing with new detectors.
- In-experiment processing was critical for single-particle analysis to understand the quality of the sample being characterised and the value of the experiment being undertaken, as well as to allow informed in-experiment decision-making.
- GPUs and high-performance-computing clusters were critical infrastructure for electron microscopy. High-end workstations continued to be relevant for specific tasks.

- Remote desktops were becoming widely adopted, both as an outcome of the Characterisation Virtual Laboratory (CVL) project and independently.
- Single-particle analysis was the most advanced and well-defined processing pipeline. Tomography-processing techniques was less mature. Materials scanning transmission electron microscopy (STEM) was less standardised as data processing and analysis are highly experiment-specific.
- Electron-microscopy software was broadly accessible on high-performance-computing clusters but users were not aware of this.

Recommendations

5. Review whether current real-time processing requirements are being met at facilities and develop a plan of support or uplift accordingly.
6. Collaborate on making data-processing software more accessible to users. There are a number of ways this could be achieved:
    - using the CVL;
    - using containerisation;
    - increased availability of services (*e.g.* CryoSPARC, LiberTEM).
7. Make data-processing infrastructure across facilities more uniform. For example, by promoting virtual desktops such as CVL with similar containerised software packages.
8. Provide users with detailed information about existing computing resources that cater to electron microscopy. For example, by developing a national repository of all the software packages together with documentation on where and how to access them. Such national repository should be promoted to users by training and targeted communication at each facility.
9. Identify opportunities to accelerate research by automating repetitive tasks. Prototype an easy-to-use solution to cater to batch jobs or pipelines for repetitive tasks.

Data management

Key findings

- Facilities provided enough storage and management for the purposes of the efficient running of the facility and for specific use cases.
- Where long-term data management was provided, it was in partnership with other (local or external) facilities such as eResearch centres or high-performance-computing facilities.
- Archiving old data off processing systems was critical to ensure the availability of sufficient storage capacity for new and ongoing processing pipelines.
- A major challenge was to understand what data can be deleted.
- There was broad interest in applying data-management tools to electron-microscopy data.
- Current practices and future plans for metadata capture appeared heterogeneous across all facilities.
- There was a large amount of information that was not harvested. Metadata capture upon data capture or upload was often missing.
- Some facilities expressed indecisiveness about capturing additional metadata

Recommendations

10. Develop a common set of best practices for data retention and data deletion.
11. Prototype and illustrate current and potential data-management software. Where a tool is already being used to manage electron-microscopy data, document this solution for the entire community.
12. Develop a national, common minimum set of persistent identifiers, metadata standards and vocabularies, as well as guidelines for metadata extraction and data transformation and the tools

and services associated with them. Build upon recommendations from the ARDC Data and Services Discovery project: “Bringing Long-Tail Microscopy and Characterisation Data into the Light”. Develop guidelines to facilitate the registration of instruments in RDA by facilities.

13. Build upon the ARDC Data and Services Discovery project: “Bringing Long-Tail Microscopy and Characterisation Data into the Light” that identified projects for data packaging that support FAIR (*e.g.* RO-Crate). A major barrier to FAIR data is the overall challenge of moving, storing and archiving EM data. The broader adoption of data-movement, metadata-capture, data-management and data-orchestration tools will provide researchers with necessary infrastructure to make EM data FAIR or FAIR-ready.

Data orchestration

Key findings

- Upon capture and processing, electron-microscopy data were described as passing through various states, including changes in data redundancy, location and ownership, which, in principle, could be automated.
- None of the institutions interviewed indicated that they were using data-orchestration tools to automatically manage states from capture to processing and then to archival.
- A method to automate the full process was desirable given the time required to manually manage the process.
- Two institutions were able to articulate a data pipeline that could be orchestrated.

Recommendation

14. Investigate and prototype orchestration tools for electron-microscopy data.

1. Introduction

1.1. Aim of the report

Over the past few years, Australian microscopy facilities have invested heavily in new-generation high-kV electron microscopy, super-resolution imaging, high-resolution block-face imaging and correlative multimodal microscopy instruments. Although recent technological developments in instrumentation have paved the way for unprecedented scientific advances, the ever-increasing amount of data produced by new technologies have posed substantial challenges to facilities in terms of how they organise and manage their workflow from the point of data capture to data storage.

The aim of the present report is two-fold:

1. To provide a review of the tools, methods and procedures used by facilities across Australia that host large-data-producing electron microscopes, including scanning electron microscopes, transmission electron microscopes, scanning transmission electron microscopes, focused ion beam systems, and derived techniques such as cryo-electron microscopy, focused ion beam–scanning transmission electron microscopy, and correlative light–electron microscopy;
2. To propose recommendations to the stakeholders of the Australian Characterisation Commons at Scale (ACCS) project, specifically under the work package "Big-data electron and correlative microscopy from instrument to publication", for tools, methods and procedures to examine further over the first year of this work package. The report will also guide all other work packages of the ACCS project into the future.

This work was undertaken under the Australian Characterisation Commons at Scale (ACCS) project, in particular under Work Package 4: "Big-data electron and correlative microscopy from instrument to publication".

1.2. Method

The information for the report was collected by interviewing representatives from academic facilities in Australia that operate, or are about to acquire, instruments that produce large volumes of data. Those facilities are hosted at institutions that are stakeholders of the ACCS. They are (see Table 1 for more details on the interviews):

- the Centre for Microscopy and Microanalysis (CMM) at The University of Queensland;
- Sydney Microscopy and Microanalysis (SMM) at The University of Sydney;
- the Electron Microscope Unit (EMU) at UNSW Sydney;
- the Cryogenic Electron Microscopy facility (referred to as UOW from here on) at the University of Wollongong;
- the Monash Ramaciotti Centre for Cryo-Electron Microscopy at Monash University;
- the Monash Centre for Electron Microscopy (MCEM) at Monash University;
- the Melbourne Advanced Microscopy Facility (MAMF) at the University of Melbourne;

- the Centre for Microscopy, Characterisation and Analysis (CMCA) at The University of Western Australia.

In addition, representatives from two facilities based in the United States and renowned internationally in the field of electron microscopy were interviewed. Those two facilities were suggested during interview. They were:

- the Simons Electron Microscopy Center (SEMC) at the New York Structural Biology Center (NYSBC);
- the Materials Research Laboratory (MRL) at University of Illinois at Urbana–Champaign.

Table 1 List of the Australian and international facilities interviewed for this report.

Facility interviewed	Interviewee(s)	Interview date
Australian facilities		
Centre for Microscopy and Microanalysis (CMM) (The University of Queensland, QLD)	Prof Roger Wepf Dr Rubbiya Ali	10.09.2020
Sydney Microscopy and Microanalysis (SMM) (The University of Sydney, NSW)	Dr Matthew Foley	09.09.2020
Electron Microscope Unit (EMU) (UNSW Sydney, NSW)	Prof Richard Tilley Dr Nicholas Ariotti Luc Betbeder-Matibet Kiho Cho Frank Li Jake Surman Dongming Zheng	25.09.2020
Cryogenic Electron Microscopy facility (University of Wollongong, NSW)	Dr Simon Brown	09.09.2020
Monash Ramaciotti Centre for Cryo-Electron Microscopy (Monash University, VIC)	Dr Georg Ramm Jay van Schyndel Hari Venugopal	07.09.2020
Monash Centre for Electron Microscopy (MCEM) (Monash University, VIC)	A/Prof Matthew Weyland Dr Peter Miller	02.10.2020
Melbourne Advanced Microscopy Facility (MAMF) (University of Melbourne, VIC)	A/Prof Eric Hanssen	15.09.2020
Centre for Microscopy, Characterisation and Analysis (CMCA) (The University of Western Australia, WA)	A/Prof Martin Saunders Dr Andrew Mehnert Dean Taylor	24.09.2020
International facilities		
Simons Electron Microscopy Center (SEMC) (New York Structural Biology Center, USA)	Dr Bridget Carragher Clint Potter	24.09.2020
Materials Research Laboratory (MRL) (University of Illinois at Urbana–Champaign, USA)	Dr Mauro Sardela Dr Timothy Spila	25.09.2020

Note, given the timeframe allocated to the report within Work Package 4, interviews of a few facilities could not be included in the present report. Those facilities were Australian national facilities (Adelaide Microscopy at The University of Adelaide, and the Centre for Advanced Microscopy and the National Laboratory for X-ray Micro-Computed Tomography, both at The Australian National University) and facilities located overseas (including European Molecular Biology Laboratory, EMBL, at Heidelberg, Germany; the Centre for Materials Elaboration and Structural Studies, CEMES, at CNRS Toulouse, France; and Maastricht University and Leiden University, Netherlands). The interviews of those facilities will be included in a later publication. We believe that this report and the information collected in all the interviews outline current and best practices in big-data electron-microscopy informatics.

Australian facilities were interviewed following a pre-established, standard set of questions in order to draw a consistent, objective and accurate picture of the current state of informatics and data management in big-data electron microscopy and correlative microscopy. The questions covered a broad range of elements related to data workflow and data management at the facilities, including: data capture, data transfer, data orchestration and overall data-workflow set-up, data-processing tools and supporting processing infrastructure, data storage, and data and metadata management.

A full set of the questions is provided in Appendix 1. Interviews with international facilities were conducted as a free-form discussion that addressed the same above points. The answers provided by Australian and international facilities were summarised, synthesised and assessed. Trends, features, tools, procedures and challenges were identified. Note, the questionnaire submitted to the facilities was not intended to be comprehensive nor exhaustive. Four major domains related to data workflow and management were considered for this report:

1. data movement;
2. data processing;
3. data management (including metadata);
4. data orchestration.

Figure 1 shows an example of the processes and tools used from data capture at the instrument to data movement and processing on different systems, then data management

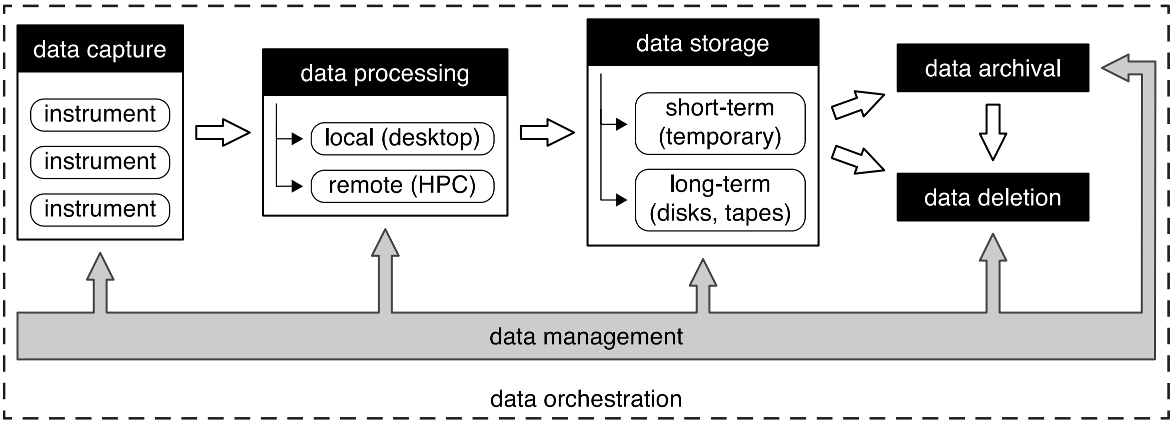

Figure 1 Example of data-orchestration workflow from the creation of data at the instrument level to data archival and deletion.

and storage. Data orchestration encompasses all the processes from data capture, to processing, archiving, data management and deletion. It considers the whole data flow from an overall perspective and how it might be optimised and automated.

The local operational environment of each facility was considered and put in perspective with the nature of the services provided to their users' communities. For example, some facilities had specific users' communities (with regard to a scientific field, an instrument and/or a technique) while other facilities offered a broader range of techniques that supported several disciplines amongst their users (*e.g.* life science and material science).

The authors of the report have outlined specific recommendations. These are explicitly documented through the report and boxed for clarity.

1.3. Scope of the report: what is big-data electron microscopy?

All the facilities interviewed were known to operate, or to be in the process of acquiring, instruments that produce large volumes of data that presented significant challenges in terms of data management and processing. It is worth noting that the notion of "big data" is relative to the environment of a facility, in particular in regard to the information-technology capabilities available and their scalability at the institution hosting the facility. This report focuses on large-data-producing electron and correlative light–electron microscopes, that is scanning electron microscopes (SEM), transmission electron microscopes (TEM), scanning transmission electron microscopes (STEM), focused ion beam systems (FIB), and derived techniques such as cryo-electron microscopy (cryo-EM), FIB-SEM and correlative light–electron microscopy (CLEM). Those techniques are used in the facilities interviewed and can generate datasets ranging from several hundreds of megabytes through to hundreds of gigabytes and tens of terabytes. Table 2 summarises the list of big-data-producing instruments at the facilities interviewed. A detailed list of the instruments is provided in Appendix 2.

2. Data movement

Data movement is the process of moving data from its point of capture or generation at an instrument to other locations for the purpose of storage, processing or archiving. This can be internal within the capture instrument, across a facility via an internal network, or between facilities via private networks or the internet. Processing includes quality control, analysis and creating and sharing datasets during collaboration.

The majority of the facilities interviewed estimated that they generated in the vicinity of half a petabyte of data annually each. Data generation per instrument ranged from 100 GB to 2–3 TB daily on some of the most data-intensive instruments. It was noted that the amount of data generated was rising quickly as data-generation and data-capture technology advanced. Big-data-producing instruments often ran a few days each week, but the experiments associated with large data volumes could extend to up to 20 days. Therefore, fast, efficient data movement off the capture device to appropriate storage device or facility was judged critical to prevent the data-capture process from stalling because of disk-space shortage in the capture device.

In this section, the network infrastructure that supports data movement at the facilities, as well as the tools and the workflows associated with data movement, are examined.

Table 2 List of the big-data-producing instruments at the facilities interviewed.

Facility[a]	Big-data producing instrument[b]	Volume of data created
CMM	1 × cryo-TEM 1 × cryo-TEM/STEM 1 × STEM 2 × SEM 1 × FIB-SEM	TEM: 500–750 TB (1 TB/day/instrument) SEM: 100 GB/day FIB-SEM: 10 GB/week STEM: 100 GB/day
SMM	3 × SEM 2 × APT 2 × XRM 5 × TEM 7 × LLM *Soon: 1 × cryo-TEM, 1 × XRM, 2 × LLM*	500 TB/year
EMU	3 × FIB-SEM 1 × cryo-TEM 1 × TEM *Soon: 1 × TEM*	Overall: 1–2 PB/year, incl.: cryo-TEM: 500 TB/year, LSFM: 500 TB/year.
UOW	2 × cryo-TEM 1 × TEM	approx. 4300 movies/24 hours, 0.5–1.5 GB/movie, up to 3 TB/24 hours
Ramaciotti Centre	2 × cryo-TEM 1 × TEM	500 TB/year (archived 09.2019–09.2020: 332.2 TB)
MCEM	1 × TEM *Soon: 1 × Ultra-high-resolution TEM*	4D-STEM and *in-situ* TEM generate big data but the total volume is hard to estimate or predict because it varies greatly with the many different workflows and acquisition conditions used.
MAMF	3 × cryo-TEM 1 × SEM	60 TB/year (⅓ raw data, ⅔ post-processed data) Cryo-TEM: 100 TB stored (raw data: 200 TB in 18 months)
CMCA	2 × XRM 1 × FIB 1 × MRI 3 × super-resolution light microscopy *Soon: 1 × cryo-TEM*	Overall: 50 TB/year

[a] APT, Atom-Probe Tomography; cryo-TEM, cryogenic Transmission Electron Microscopy; FIB, Focused Ion Beam; LLM, Light/Laser Microscopy; LSFM, Light-Sheet Fluorescence Microscopy; MRI, Magnetic resonance imaging; SEM, Scanning Electron Microscopy; STEM, Scanning Transmission Electron Microscopy; TEM, Transmission Electron Microscopy; XRM, X-Ray Microscopy.

[b] CMCA, Centre for Microscopy, Characterisation and Analysis; CMM, Centre for Microscopy and Microanalysis; EMU, Electron Microscope Unit; MAMF, Melbourne Advanced Microscopy Facility; MCEM, Monash Centre for Electron Microscopy; SMM, Sydney Microscopy and Microanalysis; UOW, Cryogenic Electron Microscopy facility at the University of Wollongong.

2.1. Network infrastructure

The ability to move data quickly and efficiently is ultimately limited by the underlying network infrastructure and how that infrastructure is tuned. Most of the Australian facilities operated at least 10 Gb/s networks internally. However, the facilities indicated that normal operations rarely exceeded 5–6 Gb/s of throughput, suggesting that further tuning and improvement of their protocols or infrastructure was possible.[1]

The network infrastructure described by most facilities (Ramaciotti Centre, EMU, MAMF, SMM and UOW) involved multiple tiers of data storage, including direct data capture at the instrument level (often supplied by the instrument vendor) through to larger on-site storage. The latter is often connected to compute for quality control and quality analysis. Data may then be moved to larger-scale storage infrastructure for collaboration or archiving.

The Ramaciotti Centre, MCEM, EMU, MAMF and UOW commented that upgrades of instruments or the delivery of new instruments were planned by the end of 2020 or over the course of 2021. These facilities pointed out that, as a result, they were expecting an increase in the total volume of data generated locally. In addition, EMU, UOW and CMM were planning to upgrade their infrastructure to increase the speed of data transfer (*e.g.* from a 10 to 100 Gb/s network). Best practice in network-infrastructure management recommends that an initial baselining of performance be conducted and then followed by a re-baseline process whenever an infrastructure component changes as this would affect the ability of data-movement tools to continue to work efficiently.

It was observed that EM facilities often did not operate their own network-infrastructure equipment but instead, used the services of organisational or third-party resources. This was often associated with a limited understanding of the network infrastructure and a lack of monitoring of its performance beyond the common question *"is the network up or down?"*. Therefore, it was noted that a clear end-to-end understanding of network infrastructure was necessary and required deeper collaboration with external partners for the facilities to benefit fully from the network deployment and configuration. Improved monitoring of the network performance would also benefit facilities by helping them understand and optimise their data-movement environments and solve network-related issues they might encounter.

Recommendation 1: Develop reproducible baselining or benchmarking of file-transfer performance, and a process for re-baselining when components of the network infrastructure change which would affect the ability of data-movement tools to continue to work efficiently. Share and distribute the information across the ACCS network.

2.2. Data-movement tools used

A range of tools commonly used were identified and categorised in four groups based on the protocol or utility they rely on:

[1] Intel Corporation. (2010). *Maximizing File Transfer Performance Using 10Gb Ethernet and Virtualization*. Accessed 12 Oct. 2020.

- SFTP protocol: FileZilla, CyberDuck, CuteFTP, RClone;
- SCP protocol: MyData, WinSCP;
- SMB protocol: RoboCopy;
- RSync utility: RClone.

The facilities often noted that they used a wide variety of software packages. They explained it by the fact that it was rarely possible to use the same tool to cover all needs, that different infrastructure or operating systems could require different software packages, and that some tools had remained in use because facility staff or users were familiar with them.

Data-movement tools fall into two broad categories with a degree of overlap:

- Existing tools used for legacy reason: In general, users know how to use them. Because they were written a number of years ago, they are configured for outdated network infrastructures (for example 10 Mb/s networks). This has resulted in inherent performance caps regardless of the underlying hardware;
- New tools: They are often based on completely new codes and optimised for 10 Gb/s (or greater) networks to provide better performance (*i.e.* take full advantage of the network infrastructure). They often require training for facilities and users.

Some of the tools cited by the facilities are interactive. Moving data is thus carried out by simply dragging and dropping a folder containing data files from one location to another. Importantly, some tools are not cross-platform, thereby leading to a mix of platforms involved in data movement (Linux, Windows and macOS). There are also underlying non-interactive data-movement tools that are used to consider, such as NFS and CIFS, which present entire data volumes as filesystems for other tools to operate on top of. Lastly, a number of tools are generic in nature so they may have in-built features that cannot be turned off. This in turn may affect performance (*e.g.* end-to-end encryption) and significantly impact overall workflow performance. Amongst all the facilities, only the Simons Electron Microscopy Center (SEMC) (USA) used higher-level data-movement tools that provided higher-level data transport.

Several facilities expressed a desire for a common platform of data-movement tools that could be used to facilitate data movement within a facility but also between facilities, analytical centres and researchers.

Some tools were cited as worth examining further. Those tools could provide better automation and data orchestration and be part of a common data-movement platform. They are:

- Globus: tool for data transfer suggested by the SEMC;
- Aspera: used at Monash University by the Monash eResearch Centre (MeRC) and the high-performance computing facility for imaging and visualisation (MASSIVE) to transfer cryo-electron-microscopy (cryo-EM) data from outside Australia;
- dCache;
- Rucio: used at and developed by CERN (European Organisation for Nuclear Research, Switzerland).

Recommendation 2: Investigate and prototype a higher-level transport service for data transport. Usability and user experience will prove important, as will long-term support.

2.3. Workflows

Data movement can be broken down into four main workflows as illustrated in Figure 2:

- Instrument to facility data storage (A in Figure 2): On-instrument storage is typically prioritised for speed so that the data can be captured quickly during measurements. However, the data need to be transferred regularly to a larger, longer-term storage facility because in-instrument storage often fills up within hours.
- Data storage to large analytical compute (B in Figure 2): While some “on-the-fly” data analysis is often performed to confirm validity, in-depth analysis often requires transfer to high-performance compute (HPC) nodes such as MASSIVE at Monash University used by the Ramaciotti Centre and UOW.
- Data storage to researchers’ individual computers (C in Figure 2): For smaller data sets, it is sometimes possible for individual researchers to run analysis themselves on a high-end workstation instead of using HPC resources.
- Archival (D in Figure 2): It is not practical nor possible to store the growing data volumes generated by instruments on the main storage servers. Therefore, once data analysis is complete and regular access to data is no longer required, the data can be archived in a long-term storage facility which can be local, remote or on the cloud.

Note, some of these workflows may also ingest data into a data-management platform.

The workflows described by the facilities were predominantly manual and required human intervention to initiate data transfer. Some facilities indicated that they had developed some form of automated scripts to move data but a high degree of manual oversight was still necessary. It is thus considered that improving the levels of automation with data-orchestration tools will help free up time and resources for other tasks.

Recommendation 3: Develop optimised and automated workflows for data movement where relevant in order to minimise human intervention for routine tasks.

Many facilities commented that, while researchers tended to focus on the data-generating end of data workflows, one of their main challenges related to data movement was managing data archival upon the completion of projects. In addition, it was observed that researchers and facilities often kept raw, processed or published data locally in case they were needed again later. However, storage servers cost money to maintain and run, and, ultimately, they all become full. So, archiving old data onto an affordable, long-term storage solution is essential and as important as generating data.

Recommendation 4: Establish reproducible processes for data archiving to ensure that computing facilities are able to continue processing new data.

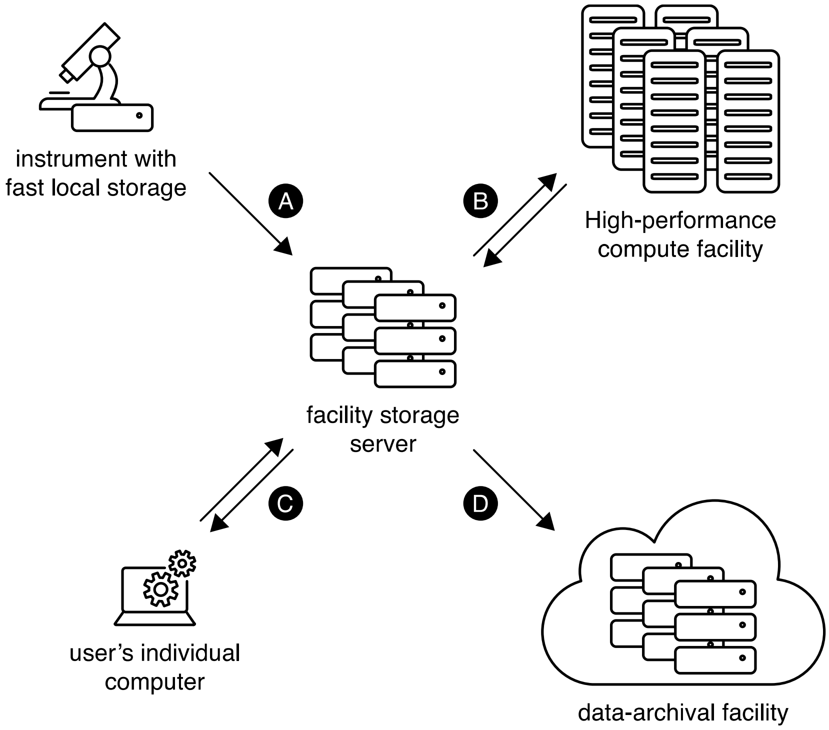

Figure 2 Schematic representation of workflows involved in data movement. A, data workflow from an instrument to a facility data storage. B, data workflow from a facility data storage to a high-performance computer. C, data workflow from a facility data storage to a user's computer. D, data workflow from a facility data storage to a data-archival facility.

3. Data processing

Raw data from electron microscopes often undergo a series of processing steps to extract the information within. Some data-processing steps can happen on-the-fly as data are generated from the microscopes. This processing is critical to understanding the quality of the sample being characterised and the value of the experiment being undertaken, as well as to allow informed in-experiment decision-making. It is considered particularly important given the expense of high-end EM instruments. The bulk of processing happens offline in other processing facilities. Many software packages perform some or most of these transformations. In order to deal with the volume and velocity of the data, data processing is often carried out in high-end workstations or high-performance computers.

Recommendation 5: Review whether current real-time processing requirements are being met at facilities and develop a plan of support or uplift accordingly.

In this section the software and the computing infrastructure used or suggested by the facilities interviewed to perform data-processing steps are presented.

3.1. Processing requirements and maturity by EM techniques

Three major EM processing techniques were discussed during the interviews with the facilities: single-particle analysis, tomography and materials scanning transmission electron microscopy (STEM) techniques. Of these, single-particle analysis showed the most advanced and best-defined processing pipeline. Tomography processing techniques seemed less mature whereas materials STEM techniques were described as highly experiment-specific so the processing steps were generally tailored to the experiment undertaken.

3.2. Data-processing software packages

A range of current and new data-processing software packages were identified. They can be categorised into four groups based on their usage:

- On-the-fly data-processing: packages capable of performing processing as data are generated from the microscopes. This type of data processing can be used for pre-processing or quality assurance of the data and is often done at workstations or computing clusters close or integrated to the instruments;
- Single-particle analysis: packages capable of performing analysis of single-molecule samples generated by cryo-EM;
- Electron tomography: packages capable of performing tomography, that is a 3D reconstruction of samples. The Ramaciotti Centre indicated that electron tomography would be used increasingly as their primary method for data processing;
- 3D rendering: packages that can be used for 3D rendering of EM data.

A detailed list of the software packages cited during the interviews is given in Appendix 3. For each package, the list details which of the four groups listed above it belongs to, the execution platform (Linux, macOS or Windows), whether it supports GPU acceleration and the nature of the software license. Note, the CVL (Characterisation Virtual Laboratory) was mentioned multiple times by many facilities as a user-friendly and powerful means to deliver software packages to users. The availability of those packages in CVL are therefore also reported.

Recommendation 6: Collaborate on making EM data-processing software more accessible to users. There are a number of ways this could be achieved:

- using the CVL;
- using containerisation;
- increased availability of services (*e.g.* CryoSPARC, LiberTEM).

3.3. Data-processing infrastructure

To better understand different aspects of the processing infrastructure employed, each of the facilities was asked to identify their processing challenges and requirements, as well as the computing infrastructures that they and their users used. Finally, facilities were invited to assess their processing-capability maturity.

The answers reflected a broad range of data-processing capabilities. This was expected because the facilities interviewed varied in terms of business models, structures and locations.

However, several common features were identified:

- Most facilities and users deployed a mixed pool of local workstations, HPC facility or clouds for data processing. Local workstations or small connected clusters were typically used for small-footprint jobs such as pre-processing for quality-assurance purposes, or on-the-fly jobs for in-experiment decision-making. More demanding jobs with higher requirements in CPU or GPU power and/or memory, such as single-particle analysis, were offloaded to remote HPC processing sites (Table 3). Note, different levels of responsibilities in data processing were described by the facilities. For example, while MCEM and SMM stated that they did not perform any processing on behalf of their users, other facilities offered some data-processing services for internal or external users, particularly in-experiment processing or value-add services;
- Virtual desktops were popular amongst facilities to cater to interactive jobs. While the CVL was available in most facilities, Virtual Research Desktop (VRD) was also available for Windows users at SMM. These platforms provided a familiar environment on powerful ready-to-use desktops for processing large datasets. Both CMM and the Ramaciotti Centre indicated that CVL was available to beginner users;
- An overall lack of support for repetitive tasks, that is batch jobs or pipelines, was noticed across many facilities. While this might not be a problem for expert HPC users, it would limit less advanced users to interactive jobs, which in turn would affect the scale and reproducibility of data analysis.

Table 3 Remote data-processing HPC sites used by the facilities interviewed.

Facility	HPC facility	HPC facility location
MAMF MCEM Ramaciotti Centre UOW	MASSIVE	Monash University (Vic)
CMM	Wiener Awoonga	The University of Queensland (Qld)
MAMF	Spartan Bio21 Cluster	University of Melbourne (Vic)
MAMF	WEHI Research Cloud	WEHI (Walter and Eliza Hall Institute of Medical Research) (Vic)
CMCA	Pawsey Cloud (Nimbus)	Pawsey Supercomputing Centre (WA)

Regarding the self-assessment of data-processing maturity, a wide range of answers was received. While MCEM, EMU and MAMF stated that their processing maturity was still at *“early stages”*, UOW judged they were at *“70% maturity”* and the Ramaciotti Centre considered their processing capability *“quite good”*. The other facilities did not provide a specific self-assessment but commented that it depended on instruments and techniques.

On data-processing challenges and associated data-processing requirements, most facilities agreed that dealing with increasingly large amounts of data from instruments was their main challenge, and thus, more infrastructure was needed. MCEM, however, considered regularisation of their workflows for certain experimental modalities and having their users to use remote HPCs such as MASSIVE were their main challenges. MCEM and SMM had difficulties articulating their processing requirements. While the former had most of their processing done locally using users' own software packages, the latter considered it depended on projects. Amongst the facilities that could describe their processing requirements, MAMF was still in the process of acquiring new workstations for new cameras, and CMM considered the CVL to be important for new users. All facilities found that access to GPUs was important to accelerate data analysis.

Recommendation 7: Make data-processing infrastructure across facilities more uniform. For example, by promoting virtual desktops such as CVL with similar containerised software packages.

Recommendation 8: Provide users with detailed information about existing computing resources that cater to EM. For example, by developing a national repository of all the software packages together with documentation on where and how to access them. Such national repository should be promoted to users by training and targeted communication at each facility.

Recommendation 9: Identify opportunities to accelerate research by automating repetitive tasks. Prototype an easy-to-use solution to cater to batch jobs or pipelines for repetitive tasks.

4. Data management

4.1. Definition

Research data management is an ensemble of practices involved in the collection, organisation, storage, preservation, documentation and sharing of research data over the course of and beyond a research project. Good practices in data management maintain the quality and the integrity of research data and ensure compliance with legal, ethical, governance and funding requirements.

An important aspect of data management is the documentation and description of data using metadata and persistent identifiers. Metadata exist in a variety of formats (*e.g.* text, HTML, XML) as separate documents, linked data files or embedded within data files. Metadata ensure that research data can be discovered, shared and reused and that experiments are reproducible. Metadata are often categorised into functional types:[2]

- descriptive metadata: information required to find and understand data (*e.g.* title, contributors, dataset descriptions);

[2] Australian National Data Service (ANDS). (2016). *ANDS Guide: Metadata.* Accessed 12 Oct. 2020. ↗

- provenance metadata: relating to the origins and processing of data (*e.g.* instrument or technology used to collect the data, processing steps);
- technical or intrinsic metadata: information required for a person or machine to read the data (*e.g.* file-level metadata such as size, checksums, mime type);
- rights and access metadata: information about data access, use and reuse (*e.g.* access rights and conditions, Creative Commons license);
- preservation metadata: relating to management of data for long-term accessibility;
- citation metadata: information required to cite the data.

A persistent identifier (*e.g.* DOIs, handles) is any label used to name something uniquely (online or offline) that is guaranteed to be managed and kept up to date over a defined period of time.[3]

The type of metadata and how and when they are harvested depend on the data model used. The ARDC Data and Services Discovery project: "Bringing long-tail microscopy and characterisation data into the light" [4] examined a range of data models and recommended a project-centric data model (Project-Subject-Study-Data, referred to as PSSD) be generalised for data management because it was considered suited to research applications. Data-management tools such as DaRIS and OMERO are already project-centric while MyTardis and 4CeeD can be mapped to a PSSD model.

4.2. Data management policies

Although all the universities that host the facilities interviewed have developed a research data management policy, the levels of awareness and knowledge of those policies and the understanding of the scope and the responsibilities that those policies may lay out (responsibilities of institutions *vs* responsibilities of researchers *vs* responsibilities of facilities) were often either minimal or uncertain to the facilities. This may have led to varying degrees of implementation of those policies across facilities. Amongst the facilities that showed an advanced level of awareness of their institutional data-handling policies and guidelines, EMU indicated that they were currently working to have the processes at the facilities compliant with university policies.

Some facilities observed that they had developed additional, local, specific data-management policies or requirements to supplement institutional policies. However, where those policies were documented and how they were implemented did not appear uniform across facilities and sometimes unclear. Those local policies or rules were often said to be described in user's manuals (*e.g.* Ramaciotti Centre and MCEM) or in the terms and conditions or registration forms for access to the facility and utilisation of instruments (*e.g.* SMM and CMCA). Those documents alongside the usage of a data-management tool (if any) were often cited as the implementation of the data-management policy at the facilities. For example, the user's manual

[3] Australian National Data Service (ANDS). (2016). *ANDS Guide: Persistent identifiers: awareness level.* Accessed 12 Oct. 2020.

[4] Wepf, R., Sullivan, R., Foley, M., Mehnert, A., Narayanan, A., Yen, L., Asomani, A., Wu, M., & Joos, A. (2019). *Bringing long-tail microscopy and characterisation data into the light.* ARDC Data and Services Discovery project.

(2019 version) at the Ramaciotti Centre (chapter "Storage Of Materials And Data, Computing") was presented as the relevant policy document for the management of data created at the facility.[5] Specifically, the user's manual states that *"[s]torage and security of [...] data [are the user's] responsibility"* and bans the use of USB memory sticks, external USB drives or similar devices on cryo-EM computers because research data *"can be lost at any time due to hardware failure, software upgrade or fault or user error"*. SMM described a different approach in which users were required to show that they had completed a research data management plan (RDMP) for their projects, or that they had a plan to transfer their data out of the facility either when they applied for access to the facility or when they wanted to retrieve their data out of SMM. SMM commented that this ensured a minimum level of data-management awareness amongst their users. The overall level of maturity in the application and awareness of data-management policies across all the facilities interviewed was best summarised by EMU by qualifying theirs as in its *"infancy"* despite their best efforts to have them aligned with the good practices recommended by funders and partners such as Microscopy Australia.

4.3. Data management responsibilities

The interviews highlighted that the responsibilities of facilities with regard to data management, in particular for data ownership and storage, varied across facilities and could also change over time at a facility. Note, the responsibilities of the facilities discussed here pertain to the responsibilities for primary instrument data but not for user-derived data. It was unclear from the interviews whether the concept of data ownership—and the responsibilities of the facilities and/or users associated with it—was understood unambiguously by everyone as the expressions "data ownership", "data property" and "data custodian/custodianship" were often used interchangeably.

Regarding data storage, all of the facilities stated that they took no responsibility for data storage. However, many facilities commented that they provided some level of storage capacity for raw data and users were often encouraged or had to use storage capabilities offered by their institutions to keep a copy of their data. For example, at The University of Western Australia, The University of Queensland, Monash University and The University of Sydney, users have virtually unlimited institutional data storage available through their institutional storage facilities (Institutional Research Data Store, Research Data Manager, Research Data Store and Store.Monash, respectively). UOW indicated that they provided storage and archival of raw instrument data (for at least six months). Although processed data (that is user-derived data) were considered as users' responsibility, UOW could provide storage space at MASSIVE through their partnership with Monash University while the data was active. However, it was noted that since approximately 50% of their users were external, the modalities of the support provided might vary, for example regarding data archival. In addition, as the environment of research data management was evolving at the University of Wollongong, how responsibilities were shared between the university, UOW and its users (internal, external) could change. CMM commented that their operational environment was transitioning from a model in which users were responsible for data storage and ownership to

[5] Monash Ramaciotti Centre for Cryo-Electron Microscopy (CryoEM). (2019). *User Manual 2019.* Accessed 12 Oct. 2020.

a new model in which the facility would become the custodian of the data through partnership.

A number of facilities provided data-management services in partnership with other institutional facilities. For example, the Ramaciotti Centre described that archival was carried out by the Monash eResearch Centre (MeRC) at Monash University, in collaboration with the facility while processed data were the user's responsibility. At the time of the interview, a total of 851.6 TB had been captured, processed and archived using this mechanism.

Many facilities cited data storage as a challenge as they were facing increasing amounts of data that required storage or archival. For example, the Ramaciotti Centre explained that although raw data were stored on MASSIVE and captured directly from the instruments, the timeframe for archiving was based on the best effort of MASSIVE to archive data, and that was a 24-hour cycle. In addition, it was noted that storage at the facility was limited, for contingencies only, with data kept at least one week before the local copy of the data was deleted. The facility handbook states that data are retained on MASSIVE for at least a month for processing by the user but, in reality, it was observed that it was kept for longer because one month was rarely enough time for a researcher to process their data. A major challenge identified by the facilities was understanding what data should be retained and what data could be deleted.

Some facilities explained that they had developed special protocols for research data storage. For example, SMM described that through the RDMP, their staff could know whether some data might be associated with ethics approval, commercial interest or confidentiality provision. As a result, for external commercial clients, SMM staff could take the necessary actions to destroy data after they had been provided to the clients. Similarly, EMU could adapt to special requirements in the case of commercial users by, for example, keeping data for an appropriate amount of time.

The two international facilities interviewed cited data management amongst their main challenges, especially data retention and data deletion. For example, MRL indicated that their users were responsible for their own data. So, when the local storage drives at the facility were filling, users' data were deleted. Users were warned when possible. SEMC described a different model in which users' data were kept for six months and then automatically deleted.

> **Recommendation 10:** Develop a common set of best practices for data retention and data deletion.

4.4. Data management tools

From the interviews, it was not possible to identify a single tool or a range of tools that was commonly used across all facilities. Facilities were often open to trialling programs to assist them in managing their data, both proprietary and non-proprietary solutions. Several proprietary tools had been tested (for example IDMS at SMM, IMS at CMM and syngo.share at EMU), and, at the time of the interviews, non-proprietary ones were being tested; specifically: 4CeeD at CMM and CMCA, OMERO at SMM and MAMF, and XNAT at EMU. Table 4 lists all the tools referred to as in current use at one or several facilities or for which there was a plan or a wish to use them. Amongst the non-proprietary programs cited, MyTardis was used by the Ramaciotti Centre, CMM, CMCA and EMU. It was described that Monash University ingested all data collected from the Ramaciotti Centre and Monash University-owned data from MAMF in the Store.Monash service (which is an instance of MyTardis). Other tools named and used by the facilities included CryoSPARC at MAMF (for personal projects), XNAT at CMCA (for MRI data) and CloudStor at EMU. Note, although CloudStor is not a data-management tool *stricto sensu*, it may be considered as a storage solution in an approach to data management. *Ad-hoc* approaches instead of specific data-management tools were indeed often preferred. That was commonly justified by limited resources for data management or specificities of the facilities. For example, SMM explained that despite trial periods for MyTardis, DaRIS and IDMS, those tools were not adopted ultimately because the range of instruments operated by SMM was too diverse for a “one size fits all” solution.

Recommendation 11: Prototype and illustrate current and potential data-management software for the big-data EM community. Where a tool is already being used to manage EM data, document this solution for the entire community.

4.5. Data formats

The facilities listed of range of file formats used, including open-source micrograph formats (*e.g.* MRC, DM3, DM4, SER, DICOM, EER), image formats (TIFF, bitmap, JPEG) and proprietary formats *(e.g.* EMI, Zeiss, Olympus, Thermo Fisher). The nature of the proprietary formats depended on the technique and the instrumentation. Facilities often indicated that proprietary formats were converted to standard, open-source file formats so data interoperability was in general not cited as a major hurdle for data management. The number of file formats used at the same facility could however be challenging. Thus, while some facilities listed a few different file formats produced (*e.g.* MRC and TIFF at the Ramaciotti Centre, and MRC, TIFF and EER at MAMF), some other facilities declared they were managing tens of formats (40–60 at SMM and 20–50 at EMU). The reason for this situation was the level of specialisation of some facilities. For example, the Ramaciotti Centre specialised in biological cryo-EM whereas SMM hosted over 160 instruments in total and was a multidisciplinary microscopy centre. In addition, most facilities commented that data-processing and data-analysis tools led to a multiplication of file formats.

4.6. Metadata

The questions on metadata during the interviews were based on the conclusions and recommendations of the ARDC Data and Services Discovery project: “Bringing long-tail microscopy and characterisation data into the light”.[4] Specifically, this project identified that metadata was crucial information required for effective long-term curation and reuse of data. Two types of metadata were considered: metadata generated by the instrument alongside the

Table 4 Data-management programs currently in use or suggested for future use for EM data at the facilities interviewed.

Program	Proprietary	Description
OMERO	no	Repository associated with the Open Microscopy Environment (OME) that allows to manage, visualise, analyse and share data. Developed at the University of Dundee (UK). Through OME and its Bio-Formats library, OMERO can read and write over 140 image file formats, both open-source and proprietary and including all major microscopy formats.
4CeeD	no	4CeeD (Capture, Curate, Coordinate, Correlate, and Distribute) is an open-source web-based platform for the management of scientific instrument data with a strong focus on materials science. Developed by the University of Illinois at Urbana–Champaign (USA). Allows to visualise, organise, curate and share data. A key element of 4CeeD is metadata capture upon data ingestion using an existing or a user-customised template. Data extractors are currently available for TEM, SEM, optical microscopy, atomic force microscopy, secondary-ion mass spectrometry and X-ray techniques.
DaRIS	no	DaRIS (Distributed and Reflective Informatics System) is an application developed by The University of Melbourne built with the Mediaflux data-management platform. Mediaflux is a proprietary data management platform developed by Arcitecta. The primary use of DaRIS is to supply data management and integration of biomedical imaging (magnetic resonance imaging and computed tomography) with instruments and research computing infrastructure (data capture, analysis, and visualisation).
MyTardis	no	Platform developed by Monash University to manage research data and metadata. Integrated with a range of scientific instruments, instrument facilities, research storage and computing infrastructure. Currently used to capture data from protein crystallography, neutron and X-ray scattering, optical microscopy, electron microscopy, medical imaging, flow cytometry, genomics and proteomics.
iRODS	no	iRODS (Integrated Rule-Oriented Data System) is an open-source data-management software developed by a consortium composed of private (*e.g.* Bayer, IBM) and public partners (*e.g.* NIH–National Institute of Environmental Health Sciences, Agriculture Victoria, Bibliothèque et Archives nationales du Québec), including universities (*e.g.* Utrecht University, University of Groningen, University of Colorado, Boulder, University College London). Many features, including virtualisation of data-storage resources, secure data sharing, customisation of metadata to all stored files.
IDMS	yes	IDMS (Integrated Database Management System) is a general data-management system that is not focused on integration with scientific instruments. Developed by CA Technologies.

Table 4 (continued)

Program	Proprietary	Description
IMS	yes	IMS (Image Management System) manages images over the entire data workflow from acquisition to documentation and includes functionalities for image visualisation, administration, annotation and analysis. Developed by Imagic. Supports a range of instruments (optical and electron microscopes, microscope cameras) and over 180 open-source and proprietary image formats (data and metadata).
syngo.share	yes	Distributed by Siemens for the management and sharing of clinical image data, multimedia data, radiological studies and clinical documents. Supports virtually all major standard file formats, including HL7, DICOM, IHE XDS and XDS-I.
CryoSPARC	yes	Developed by Structural Biotechnology Inc. Available free of charge for non-profit academic use. Primarily a tool for cryo-EM data processing but also contains modules for data management in order to manage data in self-consistent, self-contained project directories that can be exported, shared or transferred between locations. Also has features for disk-space management, data archival and retrieval, and metadata capture, creation and management.

data (embedded in data files or in separate files) and information created through human input. The project also highlighted the need for the development and delivery of services that support persistent identifiers (*e.g.* DOI, ORCID, Handle), metadata standards and vocabularies, as well as metadata-extraction and data-transformation tools or services such as the NSCA Brown Dog project. The project thus recommended that metadata be captured at the time of conceptualisation and upload, that standards for metadata and persistent identifiers be defined and adopted by the project partners (Microscopy Australia and the Australian National Imaging Facility) and that tools and services such as the NSCA Brown Dog project be explored.

The interviews depicted heterogenous practices and future plans across all facilities for metadata capture. Only metadata related to instruments were collected. Instrument metadata were collected and embedded in data or stored alongside data depending on the file formats used or the approaches taken for data acquisition. The nature of instrument metadata and the range of metadata captured and stored could vary across instruments, acquisition software and file formats within a facility and between facilities. For example, MAMF noted that while instrument-setting information was embedded in MRC files, it was lost when MRC files were exported to a TIFF format. Although the metadata could be kept in associated files, it was judged difficult to manage all the files generated by the process. Similarly, the Ramaciotti Centre cited the EPU (E Pluribus Unum) application that could generate XML files containing large amounts of metadata (such as lens settings, camera and detector), but the usage and processing of the metadata was left to the user. Many proprietary formats were said to have some form of metadata embedded (*e.g.* energies, scan time) but the conversion to standard or open-source formats (*e.g.* TIFF, BMP) often resulted in the loss of metadata. It was observed that some instruments recorded some metadata in the headers of TIFF files. Two facilities, namely CMCA and EMU, described noteworthy approaches to metadata. CMCA reported that every instrument had a persistent identifier (PID) registered in Research Data Australia (RDA) that was included in the metadata collected, and that a special schema was used for metadata storage for NIF-certified dataset post ingestion. As for EMU, they explained that metadata from instruments were captured and stored in perpetuity.

While the facilities articulated processes to capture instrument metadata, there was little mention of collection of metadata on experiments or any other additional types of metadata. Indeed, CMM and CMCA commented that there was a large amount of information that was not harvested and that the capture of metadata upon data capture or upload was often missing. However, SMM and MAMF indicated that some metadata related to experiments were available later after experiments had been conducted but it was the responsibility of the users or data owners to request such information (or any other kind of metadata relevant to given data). For example, MAMF could provide processing information such as pixel size to assist processing. The information was then shared as text files to the user in emails. As a form of additional metadata capture and storage, UOW noted that the convention used for folder names included the experiment date, time, academic lead and nature of the sample.

Some facilities expressed indecisiveness as to the capture of additional metadata as there was no systematic extraction process in place and they could be limited by their current setup (instrument, software). Further, the ability to store such metadata and the potential use of them were unclear. In contrast, other facilities were interested in capturing additional metadata. For example, EMU cited ongoing work to set up digital laboratory books to capture more metadata. MAMF also mentioned an opportunity to develop an electronic version of the laboratory book to capture metadata.

Recommendation 12: Develop a national, common minimum set of persistent identifiers, metadata standards and vocabularies, as well as guidelines for metadata extraction and data transformation and the tools and services associated with them. Build upon recommendations from the ARDC Data and Services Discovery project: "Bringing Long-Tail Microscopy and Characterisation Data into the Light". Develop guidelines to facilitate the registration of instruments in RDA by facilities.

4.7. Compliance with the FAIR data principles

Two kinds of approaches were described by the facilities with regard to aligning the data they produced with the FAIR data principles (Findable, Accessible, Interoperable, Reusable data).[6] On one hand, facilities such as SMM and MAMF expressed no plan to make the data they created FAIR. SMM argued that the decision to make data compliant with the FAIR principles fell to researchers because data were their responsibility. On the other hand, other facilities

[6] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*(1), 160018.

including the Ramaciotti Centre, CMM and CMCA decided to take active steps to make the data from their instruments FAIR or FAIR-ready. For example, for MRI at CMCA, metadata were extracted alongside the data. CMCA indicated that every instrument had a PID and a record identifier in RDA to ensure that data were findable. As CMCA and CMM were planning to trial 4CeeD, they expected to have more FAIR-compliant tools available.

Interestingly, some facilities associated the promotion of FAIR data with that of Open science although Open science and FAIR are two distinct approaches. For example, the institutional data-storage facility Store.Monash at Monash University was described as having the functionality to make datasets publicly available (by default datasets are private) but, once the datasets were archived to Store.Monash, it was the responsibility of the researcher to manage this. However, it was assumed that some—if not most—researchers were probably not aware their data were archived to Store.Monash or the functionality it offered. Similarly, EMU indicated that users were encouraged to publish through open-access journals and share their data via Open Data Bank or EMDataBank.

> **Recommendation 13:** Build upon the ARDC Data and Services Discovery project: “Bringing Long-Tail Microscopy and Characterisation Data into the Light” that identified projects for data packaging that support FAIR (*e.g.* RO-Crate). A major barrier to FAIR data is the overall challenge of moving, storing and archiving EM data. The broader adoption of data-movement, metadata-capture, data-management and data-orchestration tools will provide researchers with necessary infrastructure to make EM data FAIR or FAIR-ready.

5. Data orchestration

5.1. Definition

Data orchestration is a relatively new term for which the definition varies depending on the target audience. For big-data-producing EM instruments, the following customised definition will be taken: data orchestration is the automation of processes applied to data from the point of capture or generation, to where data are stored, how they are processed, and then managed for long-term accessibility. Data orchestration encompasses the tools used for data movement, processing, storage and management. In many ways, data orchestration is a new term describing an old problem: the management of data from collection to processing and long-term storage. The challenge has become greater as instruments are now capable of generating many terabytes of data per day.

The characteristics of data generated by the instruments considered by this report make the automation of the data workflow very appealing. Specifically:

- Data sets generated are large and therefore difficult to manage using manual methods.
- Because of the increasing data size and the associated processing requirements, active data need to be moved off processing infrastructure quickly to allow ingestion of new data.

- Much of the community applying these techniques is very new to using high-end computing infrastructure.

Orchestration has therefore the potential to increase researchers' productivity and is likely critical to future scalability.

5.2. Orchestration tools

The process of orchestrating data causes data to pass through various states, including changes in data redundancy, location and ownership. The data-flow state diagram in Figure 3 represents data states at the MASSIVE HPC facility for cryo-EM data generated at the Ramaciotti Centre and MAMF data owned by Monash University as data move from capture to storage, to processing and archival. As can be seen, from a user's perspective, the data flow and the data states explored are straightforward: data are collected from the instrument and moved to the user's local or remote storage prior to processing. Processed data and data derived from subsequent analysis are then published. Nevertheless, at the levels of the imaging or microscopy facility where the instrument is located and the associated HPC facility that provides storage and archival services, data or instances of data move between multiple

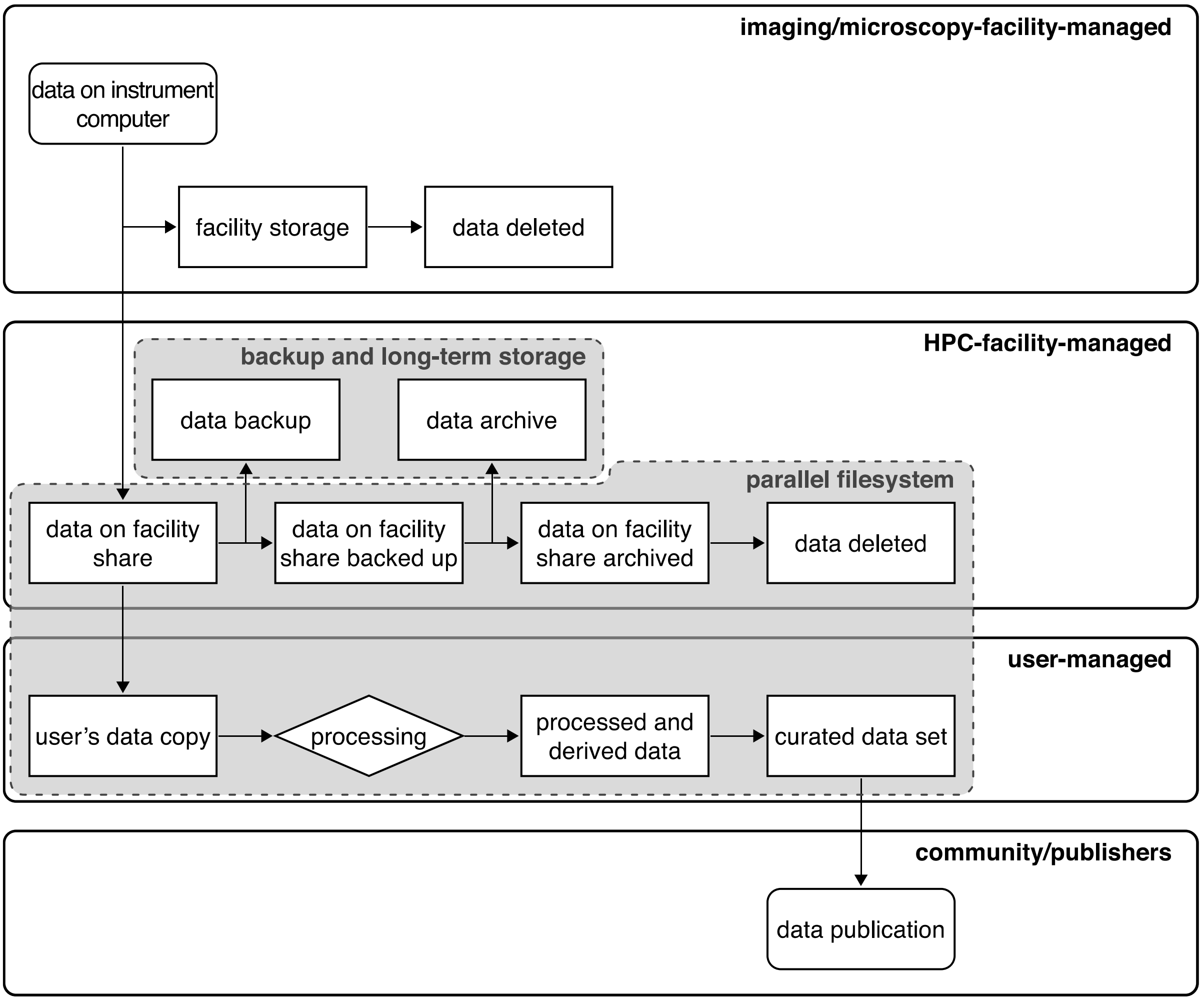

Figure 3 Data-flow state diagram for cryo-EM data at the MASSIVE HPC facility at Monash University.

states several times: short-term storage, mid-term storage and long-term storage (archival), internal transfer, transfer between facilities and deletion. The flow between those data states occurs before or after the management of the data has been transferred to the user.

For big-data-producing EM instruments, a data-orchestration system should be configurable to automatically move data from the instrument to storage for processing and then to archival. Automation is desirable given the time required to manually manage the process can be long. None of the facilities interviewed indicated that they were using data-orchestration tools to automatically manage the states that data pass through from capture to processing and then to archival. Several facilities showed some level of automation for data movement (*e.g.* RoboCopy, RSync, WinSCP, scripts, PowerShell) and processing but not for the full process. Manual tasks appeared quite common. Some of the facilities commented that they supported data capture for external parties only, not full data processing and management. On the users' side, processing steps such as structural determination are not considered to be automated as those steps were considered as highly specific to experiments and samples, and thus required significant human expertise.

A review of available data-orchestration tools suggested that there was none specific to EM. However, programs that may be appropriate were identified provided some level of configuration and adjustments could be achieved. This would require a deeper level of investigation that is beyond the present report. Table 5 lists open-source orchestration tools that have been identified as potentially interesting for big-data-producing EM instruments. Those tools focus mostly on data movement and storage. A workflow-management system might also be required to bring all the components together in a seamless orchestration process.

Recommendation 14: Investigate and prototype orchestration tools for EM data.

5.3. Data orchestration at overseas scientific organisations

In this section, the way that organisations located overseas manage the large volumes of data that they generate is briefly reviewed. Note, the data handled by those organisations are diverse and not necessarily from electron or correlative microscopy.

Materials Research Laboratory, University of Illinois at Urbana–Champaign (USA)

The Materials Research Laboratory is a large, shared facility. Access is available to all universities and businesses. Users are responsible for their own data. Data are processed on laboratory workstations or using the researcher's own processing system. Automatic data orchestration is not implemented.

New York Structural Biology Center (USA)

The New York Structural Biology Center (NYSBC) consists of nine academic research institutions and provides resources to its members and outside users. Data are kept for six months before being automatically deleted. Pre-processing and processing infrastructure are provided to users.

Table 5 Open-source orchestration-tools identified as of potential interest for big-data-producing electron microscopy.

Program	Description
Rucio	Scientific data-management software developed by CERN (European Organisation for Nuclear Research) to meet the requirements of the high-energy physics experiment ATLAS at the Large Hadron Collider. ATLAS uses Rucio as the principal distributed data-management system. The workflow-management system PanDA and the task definition and control system ProdSys are also used to bring together a complete data-orchestration system (Barisits, M., Beermann, T., Berghaus, F., Bockelman, B., Bogado, J., Cameron, D., Christidis, D., Ciangottini, D., Dimitrov, G., Elsing, M., Garonne, V., di Girolamo, A., Goossens, L., Guan, W., Guenther, J., Javurek, T., Kuhn, D., Lassnig, M., Lopez, F., … Wegner, T. (2019). Rucio: Scientific Data Management. *Computing and Software for Big Science, 3*, 11.).
Apache Airflow	Program to develop workflows. Workflows are defined in Python code and monitored in a Web UI. If used to cover the data states in EM data orchestration, Airflow would rely on external tools to perform data movements, processing, backup and archiving.
Apache Airavata	Framework that supports execution and management of computational scientific applications and workflow in grid-based systems. The main focus is on submitting and managing applications and workflows. In an EM data-orchestration system, Apache Airavata would join all the steps of data movement and processing together.
iRODS	Data-management system that contains a rules engine for workflow automation where events can trigger rules to execute. The rules can be configured to perform tasks (*e.g.* metadata extraction, data movement). Data are typically accessed through the iRODS client.
Galaxy Project	Web-based system for data-intensive biomedical research. Contains a workflow system that may be useful in orchestrating EM data.
Apache Taverna	Workflow-management system with a focus on scientific workflows. The system consists of both command line and a graphical user interface. It has been built to operate over a wide range of domains.
Next Generation Archive System (NGAS)	Developed by the European Southern Observatory (ESO) for the general handling of data (management, transport *etc*). Currently handles hundreds of millions of files and tens of petabytes of astronomical data.

European Organisation for Nuclear Research (Switzerland)

At the European Organisation for Nuclear Research (CERN), the ATLAS experiment of the Large Hadron Collider uses Rucio, PanDA and ProdSys for data orchestration. Rucio is used to transfer data around the world for storage and processing. PanDA is a workflow-management system and ProdSys allows task definition and control.

Howard Hughes Medical Institute, Janelia Research Campus (USA)

Users are requested to provide access to their network drives for uploading data. Alternatively, they need to send in large portable hard drives for data capture and shipment.

European Molecular Biology Laboratory, Heidelberg (Germany)

External users are encouraged to bring portable hard drives for data capture and storage.

5.4. Example of workflows at Australian facilities

The data workflows at the Centre for Microscopy and Microanalysis (CMM) at The University of Queensland and the HPC facility MASSIVE at Monash University are described and briefly discussed. Those workflows focus on data generated by cryo-EM instruments.

Centre for Microscopy and Microanalysis (The University of Queensland)

The diagram in Figure 4 illustrates the workflow for cryo-TEM data from data capture to

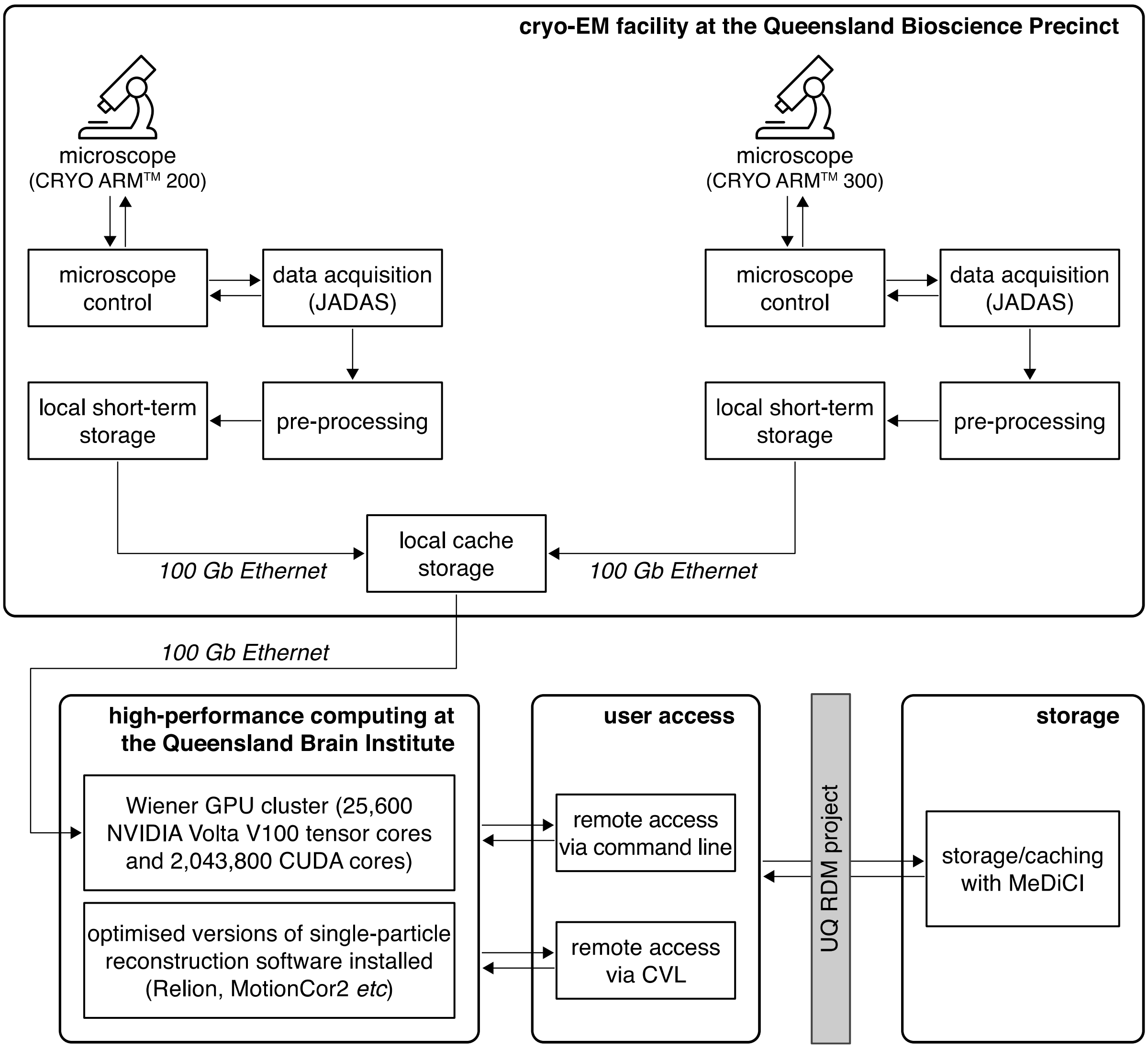
cryo-EM facility at the Queensland Bioscience Precinct
microscope
(CRYO ARM™ 200)
microscope control
data acquisition (JADAS)
pre-processing
local short-term storage
microscope
(CRYO ARM™ 300)
microscope control
data acquisition (JADAS)
pre-processing
local short-term storage
100 Gb Ethernet
local cache storage
100 Gb Ethernet
100 Gb Ethernet
high-performance computing at the Queensland Brain Institute
Wiener GPU cluster (25,600 NVIDIA Volta V100 tensor cores and 2,043,800 CUDA cores)
optimised versions of single-particle reconstruction software installed (Relion, MotionCor2 *etc*)
user access
remote access via command line
remote access via CVL
UQ RDM project
storage
storage/caching with MeDiCI

Figure 4 Cryo-TEM data pipeline for single-particle data acquisition at the Centre for Microscopy and Microanalysis at The University of Queensland (JADAS, JEOL Automated Data Acquisition System; MeDiCI, Metropolitan Data Caching Infrastructure; RDM, Research Data Manager).

processing and management at CMM at The University of Queensland. The pipeline for single-particle data acquisition in cryo-TEM is similar to that at other institutions and can thus be regarded as a typical data orchestration for big-data-producing EM instruments.

At the stage of capture, data are acquired from the two cryo-EM instruments using the proprietary software JADAS (JEOL Automated Data Acquisition System). Data pre-processing is then performed at the instrument level before the data are moved to a local cache for temporary storage. From there, the data are accessed for processing on a HPC cluster (Wiener) that consist of GPUs. The HPC infrastructure provides users with access to data from command line and the local CVL. After processing, data are stored on the university's long-term storage and archival facility called Research Data Manager (UQ RDM). Data are transferred to a specific project on UQ RDM which can be created *ad hoc* for the sole purpose of storing the experimental data or can be an existing user's project. Note, UQ RDM is indeed the interface to the physical high-performance data storage facility developed by the university called MeDiCI (Metropolitan Data Caching Infrastructure). MeDiCI provides both long-term data storage and temporary data storage for caching. Caching allows fast access to data for researchers via UQ RDM. When data have not been accessed for a while, they are automatically moved from the cache to long-term storage.

Monash University

At Monash University, the HPC facility MASSIVE provides services for cryo-EM data to both internal and external sources. Internally, MASSIVE handles cryo-EM data from the Ramaciotti Centre for Cryo-Electron Microscopy and the Monash Institute of Pharmaceutical Sciences (MIPS). MASSIVE is responsible for the whole orchestration of their data, that is movement, processing, management and storage. In addition, MASSIVE is in charge of processing cryo-EM data from the Cryogenic Electron Microscopy facility at the University of Wollongong (UOW). Finally, MASSIVE provides data-processing services to international facilities through research collaborative agreements with Monash University.

Figure 5 shows the details of the data workflow at MASSIVE for cryo-EM data using the CPU/GPU cluster named M3. The workflow is split into four steps: 1. data transfer to MASSIVE; 2. allocation of data to projects and temporary data storage; 3. data processing and analysis; and 4. data archival. MASSIVE carries out all four steps for internal users, that is the Ramaciotti Centre and MIPS, whereas UOW and international facilities are covered by steps 1–3, the data being transferring back to the institutions. UOW manage their own long-term data archival separately. Data transfer within Monash University uses either the university-wide network for MIPS or the Monash eResearch Network (MeRN) for the Ramaciotti Centre. External domestic users (UOW) rely on the AARNet network. Data transfer from and to international facilities is based on international network capabilities and AARNet. Data from international facilities can be transferred regularly from some institutions or intermittently from others. In both cases, data are transferred using the software packages Globus or Aspera. Once the data are transferred to MASSIVE, they are stored in dedicated projects depending on their origin for a set period of time. Those project-based directories are used as temporary data storage while data processing and analysis are performed. A range of HPC hardware and software solutions are

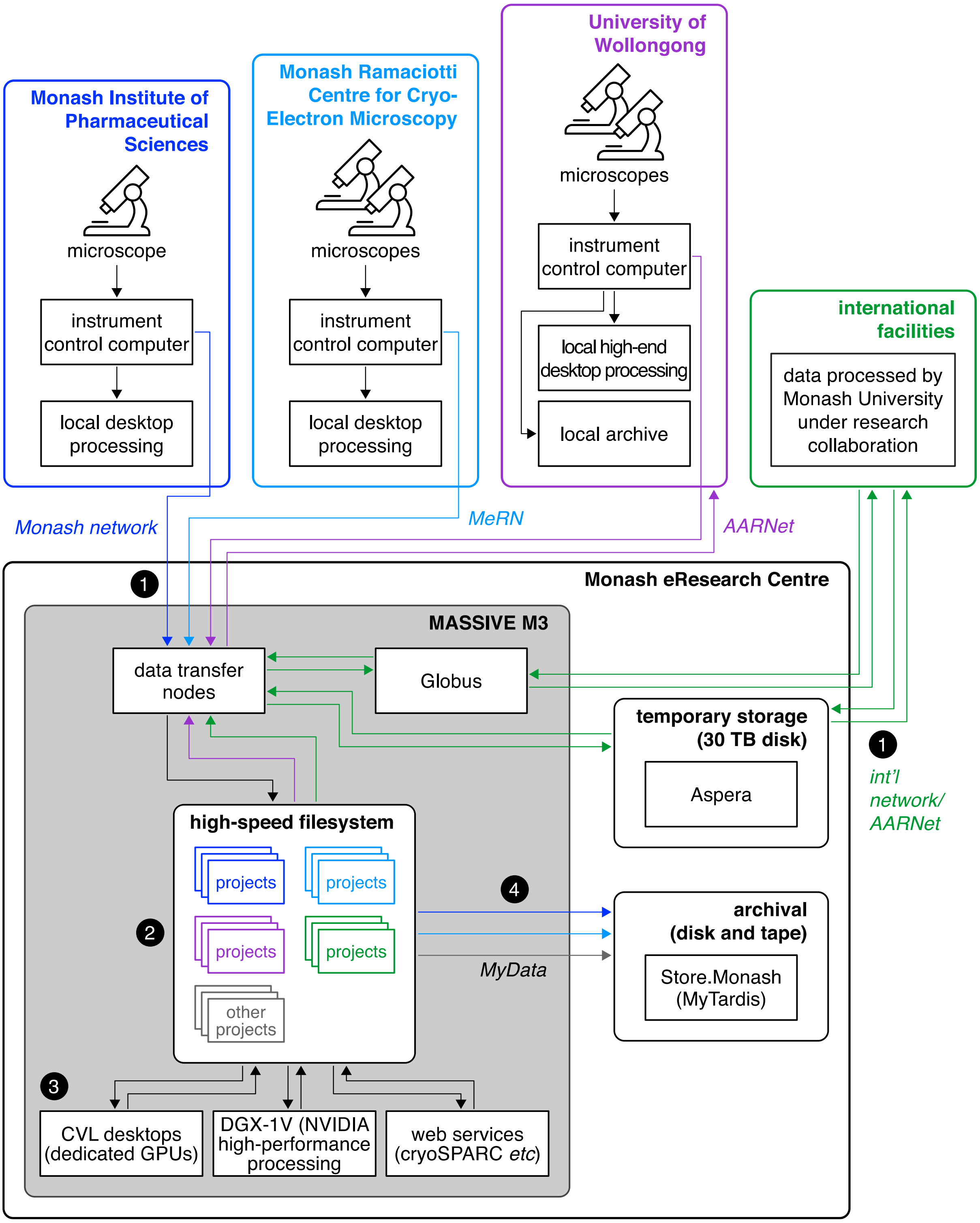
Monash Institute of Pharmaceutical Sciences
microscope
instrument control computer
local desktop processing
Monash Ramaciotti Centre for Cryo-Electron Microscopy
microscopes
instrument control computer
local desktop processing
University of Wollongong
microscopes
instrument control computer
local high-end desktop processing
local archive
international facilities
data processed by Monash University under research collaboration
Monash network
MeRN
AARNet
Monash eResearch Centre
1
MASSIVE M3
data transfer nodes
Globus
temporary storage (30 TB disk)
Aspera
1
int'l network/ AARNet
high-speed filesystem
projects
projects
projects
projects
other projects
2
4
MyData
archival (disk and tape)
Store.Monash (MyTardis)
3
CVL desktops (dedicated GPUs)
DGX-1V (NVIDIA high-performance processing
web services (cryoSPARC etc)

Figure 5 Cryo-EM data workflow at MASSIVE at Monash University (MeRN; Monash eResearch Network).

available at M3. Similarly to the data workflow described previously for CMM (Figure 4), a local CVL is available. The CVL rests upon three different types of hardware depending on how resource-intensive (memory, CPU, GPU) data-processing and data-analysis tasks are. Those are NVIDIA Tesla K80, Tesla P100 and Tesla V100 GPU cards. For other tasks not supported by or not suitable for CVL, such as data processing using Relion and MotionCor2, the specialised high-performance processing server NVIDIA DGX-1V composed of CPUs and GPUs is available. Finally, some web services such as CryoSPARC are also accessible to users. After processing, internal users' data are archived after a set period of time using the desktop application MyData for uploading data to the data-management system MyTardis. Alternatively, internal users can move their data to their temporary project allocation.

The cryo-EM data-archival workflow is currently a manual process in need of automation and would be a suitable test case for a data-orchestration tool. Note, the steps are identical for both the Ramaciotti Centre and MIPS. A detailed description of the manual process is provided in Appendix 4. A data-orchestration tool should be capable of:

- Detecting new datasets and reporting on them;
- Interacting with third-party tools to commence data movement, archiving and verification;
- Configurable workflow;
- Configurable rules; for example, archive data a set number of days after creation, verify an archived dataset and move data to required location;
- Providing visibility of dataset status; for example, data-movement status (started, completed, running, error), archiving (started, completed, error), dataset verification (started, completed, error), dataset safe to delete (*i.e.* fully archived);
- Handling sensitive or private datasets in a secure manner;
- Initiating data processing; for example, pre-processing of captured data.

6. Conclusion

The interviews of representatives from a range of microscopy facilities in Australia and overseas that operate, or are planning to operate, electron microscopy (EM) and correlative light–electron microscopy instruments that produce large volumes of data, have provided valuable information on the current landscape in informatics and data management in the field. Although this report is not meant to be systematic or exhaustive, it has allowed to identify general trends, tools, procedures, gaps and challenges across all or a majority of the facilities. Those were summarised and analysed around four key axes: data movement, data processing, data management (including data documentation using metadata) and data orchestration.

Most of the facilities interviewed declared that they relied on their institutional network infrastructure. As a result, it was observed that there could be limited knowledge or understanding of the network configuration and capabilities. It was concluded that improving network monitoring and developing reproducible baselining or benchmarking of file transfer performance would contribute to optimising data movement. In addition, optimised and automated workflows for data movement as well as processes for data archiving would benefit facilities and their users.

Regarding data processing, two aspects were covered: processing software packages and processing infrastructure. A detailed list of current and future EM-processing software packages named by all facilities was compiled. Making this list accessible would benefit all users. The processing infrastructure across all the facilities showed great diversity, which reflected a wide range of processing capabilities. In order to facilitate data processing for users, it is recommended to make the processing infrastructure more uniform across facilities.

The interviews highlighted that all the facilities had policies or guidelines for the management of primary, raw instrument data (as opposed to user-derived data) but how they were implemented, to what extent additional policies developed by the facilities (if any) supplemented institutional policies and how they were understood by users was sometimes unclear. The facilities often commented that they did not own the data. As a result, no responsibility for data storage was taken. This was all the more important that many facilities were dealing with increasing amounts of data. However, many facilities described that they provided some forms of storage for raw data but there was no mid or long-term guarantee offered to the users. Developing a common set of best practices for data retention and deletion would help the facilities know what data to store and for how long. A broad range of tools for data management was cited but there is a lack of shared knowledge across facilities. The wish and the ability to capture metadata, use persistent identifiers and produce data compliant with the FAIR principles were diverse across all facilities. It is recommended to develop standards and guidelines that will promote and facilitate the extraction and use of metadata and persistent identifiers, which will in turn support the adoption of the FAIR principles.

None of the facilities interviewed had a complete data-orchestration system implemented from capture to long-term storage and archival. Despite the absence of tools with a primary focus on data orchestration specific to the EM community, other tools available that may fit with some needs were identified. However, given that there is no tool that can cover all requirements (*e.g.* processing, movement, management, workflow), a combination of tools could provide a complete solution. An in-depth trial of a few tools is therefore recommended. The growing number of big-data-producing instruments requires that data orchestration be automated in order to manage ever-increasing data volumes properly.

Appendix 1

Standard questionnaire used for the interviews of Australian facilities

A. Instruments

A1. What big-data instruments (and detectors on those instruments) does your facility have?

B. Data capture and data transfer

B1. What data-transfer technologies are you using (and for what application(s))?

B2. Do you have a network map or network description to support the transfer of instrument data to data processing?

B3. What are the average data-transfer speeds you are currently getting (1Gb/s, 10Gb/s)?

B4. What type of large data volume have you dealt with over the last 12 months? For example, fast readout of multiple files or cameras, multiple images for 3D volume imaging, multiple diffraction patterns at very high frame rate.

B5. What are you aiming for in the next 12 months and how do you plan to get there?

C. Data orchestration and overall data movement workflow

C1. Is there any automated data movement or data processing being applied at your facility?

C2. Can you draw the current data workflow from experiment to analysis to archive or deletion or publication?

D. Processing tools

D1. What processing tools are you and your users using?

D2. Are you using (near-)real-time processing for your instrument?

D3. Do you have any new tools you would like to see supported in the future?

D4. What are your processing challenges?

E. Processing infrastructure

E1. Where is the facility processing its data?

E2. Where are the facility users processing their data?

E3. Can you articulate your processing requirements (for example GPU/CPU power, GPU/CPU hours)?

E4. Can you assess how mature your facility processing capability is?

E5. Can you name other reference sites that you think have great or advanced processing infrastructure?

F. Data management (including repositories)

F1. Is there a data-management policy at your institution?

F2. Is there a data-management policy at your facility?

F3. What is your responsibility with regards to data management, for example data ownership and data storage?

F4. If you are applying a data-management policy, how is it implemented at your facility?

F5. What data-management tools are you using for big-data EM (for example, MyTardis)?

F6. What data-management tools have you heard of? Or would like to try?

F7. Do you do anything to make the data your facility produces more FAIR (Findable, Accessible, Interoperable, Reusable)?

G. Data and metadata

G1. What data do(es) your instrument(s) typically output, or what type of data do you hand over to users?

G2. If the output data format is a proprietary standard, could you provide more information about interoperability?

G3. Do you have to deal with many different types of file formats? If yes, how many (approximately)?

G4. What (if any) metadata are you capturing with experiment? Specifically:

- Experiment metadata: Do your users or your facility add extra metadata related to the experiment or research project?
- Instrument metadata: does your instrument tag your raw data with metadata?
- Other metadata?

G5. Where do you store those metadata?

G6. Do you have a desire to capture additional metadata?

G7. Does the instrument software allow you to supply additional metadata (for example, sample)?

H. Data storage

H1. Is there a clear understanding of the data storage volumes for your instrument(s)?

- how much data produced over the last X period?
- how much projected over the next 1, 3 and 5 years?
- How much data per experiment? How many experiments per day or week?

H2. What is the storage capacity of your facility?

H3. Does your storage capacity keep up with the data being generated?

H4. Do you have a data-storage model with hierarchies (for example, capture storage, processing storage, archival storage, deletion)?

I. Overall

I1. What are the top-3 challenges you are facing?

I2. What are your training requirements?

I3. Could you please suggest an exemplar user we should talk to?

Appendix 2

Detailed list of big-data-producing instruments at the Australian facilities interviewed (as of October 2020)

Facility	Big-data-producing instrument[a]
Centre for Microscopy and Microanalysis (CMM) (The University of Queensland, QLD)	• cryo-TEM: ○ JEOL JEM-Z200FSC • cryo-TEM/STEM: ○ JEOL JEM-Z300FSC • STEM: ○ Hitachi HF5000 • SEM: ○ Thermo Scientific Apreo ○ Zeiss 3View • FIB-SEM: ○ Thermo Scientific Scios
Sydney Microscopy and Microanalysis (SMM) (The University of Sydney, NSW)	• SEM: ○ Zeiss Sigma VP 3view ○ Zeiss Sigma VP HD ○ Zeiss UltraPlus • APT: ○ Cameca LEAP 3000 Si ○ Cameca LEAP 4000X Si • XRM: ○ Bruker SKYSCAN 1272 ○ ZEISS Xradia MicroXCT-400 • TEM: ○ FEI Tecnai T12 ○ FEI Themis-Z ○ JEOL JEM-1400 ○ JEOL JEM-2100 ○ JEOL JEM-2200FS • Light optical and laser microscopy: ○ Leica Aperio XT ○ Leica SP8 DIVE ○ Leica TCS SP8 STED 3X ○ Nikon Ti-E ○ Nikon A1R ○ Nikon C2 ○ Olympus VS120
Electron Microscope Unit (EMU) (UNSW Sydney, NSW)	• FIB-SEM: ○ Zeiss AURIGA ○ Thermo Scientific Helios G4 PFIB UXe

	DualBeam ○ FEI xT Nova NanoLab 200 • cryo-TEM: ○ Thermo Scientific Talos Arctica • TEM: ○ JEOL JEM-F200
Cryogenic Electron Microscopy facility (University of Wollongong, NSW)	• cryo-TEM: ○ Thermo Scientific Titan Krios ○ Thermo Scientific Talos Arctica • TEM: ○ Tecnai T-12
Monash Ramaciotti Centre for Cryo-Electron Microscopy (Monash University, Vic)	• cryo-TEM: ○ Thermo Scientific Titan Krios ○ Thermo Scientific Talos Arctica • TEM: ○ Thermo Scientific Helios G4
Monash Centre for Electron Microscopy (MCEM) (Monash University, Vic)	• TEM: ○ FEI Titan3 80-300
Melbourne Advanced Microscopy Facility (MAMF) (University of Melbourne, Vic)	• cryo-TEM: ○ Thermo Scientific Titan Krios ○ Thermo Scientific Talos Arctica ○ Thermo Scientific Glacios • SEM: ○ Thermo Scientific Teneo
Centre for Microscopy, Characterisation and Analysis (CMCA) (The University of Western Australia, WA)	• XRM: ○ Zeiss Versa 520 ○ Skyscan 1176 • FIB: ○ FEI Helios Nanolab G3 • MRI: ○ Bruker BioSpec 9.4T • Super-resolution light microscopy: ○ Nikon A1R (PicoQuant FCS/FLIM unit) ○ Nikon Eclipse Ti2 ○ Nikon STORM

[a] APT, Atom-Probe Tomography; cryo-TEM, cryogenic Transmission Electron Microscopy; FIB, Focused Ion Beam; MRI, Magnetic resonance imaging; SEM, Scanning Electron Microscopy; STEM, Scanning Transmission Electron Microscopy; TEM, Transmission Electron Microscopy; XRM, X-Ray Microscopy.

Appendix 3

List of data-processing tools in electron microscopy cited during the interviews

Name	Description	Category	Platform	CPU/GPU	License	Availability in CVL		
						Desktop	Container registry	Command line
CryoSPARC	Processing of cryo-EM single-particle data.	on-the-fly single-particle analysis	Linux (workstations, HPC)	both	free for academic	yes	no	yes
Relion	Stand-alone program that employs an empirical Bayesian approach to refinement of (multiple) 3D reconstructions or 2D-class averages in cryo-EM.	on-the-fly single-particle analysis	Linux (workstations, HPC)	both	GPL v2	yes	yes	yes
Scipion	Image processing to obtain 3D models of macromolecular complexes using EM.	on-the-fly single-particle analysis	Linux (workstations, HPC)	both	GPL v2	yes	no	yes
Cistem	Processing of cryo-EM images of macromolecular complexes to obtain high-resolution 3D reconstructions.	on-the-fly single-particle analysis	Linux (workstations, HPC)	CPU	Janelia Open-Source Software	no	yes	yes
MotionCor2	Correction of beam-induced sample motion recorded on dose-fractionated movie stacks.	on-the-fly	workstations, HPC	GPU	free for academic	no	yes (in the Relion container)	yes

Name	Description	Category	Platform	CPU/GPU	License	Availability in CVL		
						Desktop	Container registry	Command line
gctf	Contrast transfer function determination, refinement and correction.	on-the-fly	Linux	GPU	free (not clear)	no	yes (in the Relion container)	yes
ctffind (v4)	Estimation of objective lens defocus parameters from transmission electron micrographs.	on-the-fly	Linux	CPU	Janelia Open-Source	no	yes (in the Relion container)	yes
WARP	Automation of all pre-processing steps of cryo-EM data acquisition and enables real-time evaluation. Correction of micrographs for global and local motion, estimation of local defocus and monitoring of key parameters for each recorded micrograph or tomographic tilt series in real time.	on-the-fly	Windows	GPU	GPL v3	no	no	no
CCP4	Macromolecular X-ray crystallography.	on-the-fly	Linux, macOS, Windows	CPU	free for academic	yes	yes	yes
Simple	Analysis of cryo-EM movies of single particles.	on-the-fly	Linux, macOS	CPU	GPL v3	no	no	yes
Frealign	High-resolution refinement of 3D reconstructions from cryo-EM images of single particles.	single-particle analysis	Linux, macOS	both	Janelia Open-Source Software	no	no	no
IMOD/PEET	Aligning and averaging particles in 3D sub-volumes extracted from tomograms.	cryotomography	Linux, macOS, Windows	CPU	free	yes (part of IMOD)	yes (part of IMOD)	yes (part of IMOD)
emClarity	Iterative tomographic tilt-series refinement that uses sub-tomograms as fiducial markers and a 3D-sampling-function-compensated, multi-scale principal component analysis classification method.	cryotomography	Linux	GPU	GPL v3	no	no	yes

Name	Description	Category	Platform	CPU/GPU	License	Availability in CVL		
						Desktop	Container registry	Command line
Dynamo	Subtomogram averaging of cryo-EM data.	cryotomography	Linux, macOS, Windows	both	free	no	no	yes
Appion-Protomo	Appion is a pipeline for processing and analysis of EM images. Protomo is for marker-free alignment and 3D reconstruction of tilt series in electron tomography.	cryotomography	Linux, macOS, Windows (Docker-based)	CPU (unclear)	Appion: Apache v2 Protomo: GPL v3	no	yes (Protomo only)	yes (Protomo only)
iLastik	Interactive image classification, segmentation and analysis.	3D rendering	Linux, macOS, Windows	both	GPL v2	yes	yes	yes
Blender	3D Computer-graphics software toolset for creating animated films, visual effects, art, 3D printed models, motion graphics, interactive 3D applications, virtual reality and computer games.	3D rendering	Unix-like (incl. Linux and macOS), Windows	both	GPL v2	yes	no	yes
Microscopy Image Browser	High-performance Matlab-based software for advanced image processing, segmentation and visualisation of multi-dimensional (2D–4D) light and electron	3D rendering	Linux, macOS, Windows (Matlab-based or stand-alone)	CPU	GPL v2	no	no	no
ImageJ/FIJI	Image processing for scientific multidimensional images.	3D rendering	Linux, macOS, Windows	CPU (some plugins may use GPU)	GPL v2	yes	yes	yes
Amira-AVIZO	Advanced 3D analysis software.	3D rendering	Linux, macOS, Windows	both	proprietary (Thermo Scientific)	yes	no	yes
Bruker 3D.SUITE	Reconstruction, inspection, visualisation, and analysis of structures. Supplied with instruments.	3D rendering, tomography	Windows	both	proprietary (Bruker)	no	no	no

Name	Description	Category	Platform	CPU/GPU	License	Availability in CVL		
						Desktop	Container registry	Command line
Oxford AZtec	A suite of software for a SEM/TEM/FIB/ EDS/EBSD particle analysis. Note, particles here are individual grains in metals or rocks that can be separated due to contrast so it is distinct from single-particle EM. Specifically optimised for high-speed throughput. Supplied with instruments.	on-the-fly elemental mapping, 2D rendering	Windows	unclear	proprietary (Oxford)	no	no	no
LiberTEM	High-throughput distributed processing of large-scale binary data sets using a simplified MapReduce programming model.	ptychography, holography reconstruction, analysis methods for all applications of pixelated STEM and other large-scale detector data.	Linux, macOS, Windows	both	GPL v3	no	yes	yes
Gatan Microscopy Suite	General imaging software for electron microscope experimental control and analysis. Supplied with instruments.	on-the-fly, single-particle analysis, 3D imaging, tomography	Windows	unclear	proprietary (Gatan)	no	no (free version)	no
Cameca IVAS	Integrated visualisation and analysis. Supplied with instruments.	tomography	Windows	unclear	proprietary (Cameca)	no	no	no

Appendix 4

Cryo-EM data-archival workflow at MASSIVE (Monash University)

Below is the description of the steps that cover the workflow for cryo-EM data archival at MASSIVE. This is currently a manual process employed for cryo-EM data created at the Ramaciotti Centre and the Monash Institute of Pharmaceutical Sciences (MIPS). Further details are available on the Store.Monash webpage and the MyTardis webpage. The archiving workflow as currently used on MASSIVE consists of separate scripts, Python code and a spreadsheet. There is a newer version of MyData (MyTardis archiving client) for the command line. Using Python, it would be possible to create a system to automate the archiving process and track the status of datasets. This would cover off a large part of the data-orchestration process. However, data capture and processing would also need to be considered.

The capture of end-user requirements would be suitable for a separate document and was not covered in the survey. Briefly, the requirements would cover the ability to track the status of a dataset (*i.e.* new, archived, processed, deleted).

<u>Data-archival workflow:</u>

1. On MASSIVE, the folders containing cryo-EM data from the Ramaciotti Centre and MIPS are checked regularly for new datasets by running a bash script. This script returns the dataset and its size.

2. A tracking spreadsheet is used to record the newly identified datasets and their sizes:
    - It is shared with each facility manager.
    - There is one tracking spreadsheet per facility containing one worksheet per instrument.

3. For new datasets, MyData, the client for MyTardis, is run on the HPC system data-transfer nodes to archive raw primary data to The Vault (MASSIVE tape system for long-term data storage). MyData returns the total number of files for a dataset. This is added to the tracking spreadsheet along with the URL from MyTardis to uniquely identify the archived dataset.

4. Once a dataset has been fully archived by MyData, a separate Jupyter Notebook (Python) script is executed to interrogate the MyData API. This script checks the status of every file in a dataset to ensure it has been archived. A total number of files, the total size of the dataset and the last archived date are returned. Then:
    - The total number of files and the total size of the dataset are compared against those in the tracking spreadsheet.
    - Any files that are not verified are identified.
    - For a dataset, the latest date for an archived file is recorded in the tracking spreadsheet. This is used to gather statistics over a period of time.

- A dataset is not considered fully archived until all files have the status `verified=True`, the file count matches and the total size of the datset is very close to that recorded from the bash script. Note, because of differences in the filesystems used across systems and the tools used to check disk usage, the reported total dataset size can vary.

5. Once a dataset has been fully verified and file totals match, the tracking spreadsheet is updated to indicate this. This is a flag to the facility manager that raw data have been safely archived and can be deleted from MASSIVE when required.

6. This process allows for the raw data to be archived while being processed by the researcher on MASSIVE.

7. If required, the data can be easily retrieved from the tape system and restored back to MASSIVE.

8. A few rules allow some level of automation. Specifically:
   - The folder structure is standardised and follows the template: `/projects/projectID/instrument/rawdata/emailAddress/yyyymmdd`.
   - This structure allows MyTardis to automatically assign ownership of a dataset by matching the email address with the user.
   - Datasets are created using the format `yyyymmdd`.
   - Once a dataset is created, it is never modified again. This ensures that files are never created after archiving has completed.

9. MASSIVE pushes cryo-EM data to Store.Monash which is the Monash-wide instance of MyTardis.