

Data Management, Processing and Modelling Working group

The management, analysis and simulation of large datasets has become an increasingly large part of the electron microscopy ecosystem as techniques are developed that generate large volumes of data. There was no specific mention of electron microscopy data in the road mapping exercise that reported in 2016, with a clear focus on hardware developments and training. Since this last exercise, it has become increasingly clear that handling of data is a major challenge for many EM facilities and that support and resources in this area should be a future priority for the community.

A survey was conducted at IOP's EMAG AGM at MMC 2023, with specific questions on data management/analysis/simulation provision in labs. The survey was completed by 21 EMAG members. Survey answers revealed that approximately 50% of the EM facilities included in the survey provide support for data management and analysis, highlighting that these aspects of research are often handled at an individual group level rather than at facility level. This may present a significant barrier to non-expert microscopy users whose research groups do not contain expertise on the handling of data.

The growth of 4DSTEM and in-situ methodologies put a particular strain on data storage and management, where the collection of TBs (or greater) of data in single experiments is standard practice. Survey results indicated that provision of data storage at EM facilities was very mixed (only 43% of facilities provided any data storage), with the majority of data storage noted to be managed through individual research groups (and funded through individual grants). There is a clear need for the community to provide guidelines for the storage of data, with many facilities reporting only local (at risk) storage of data while others utilize fully protected cloud storage. Guidelines for the management of data would ensure data security and allow for researchers to fully cost the necessary provision into research grants.

The future management of data will not simply be a question of where data is stored. UKRI have highlighted that one of their priorities is that all research data collected is FAIR (findable, accessible, interoperable and re-usable). Currently, the majority of data published in physical sciences electron microscopy is solely available on request from manuscript authors. In the field of cryo-EM there have been huge strides towards FAIR data. Publicly accessible data archives exist for cryo-EM/ET (EMPIAR and EMDDB) in which all published data is searchable and downloadable by anyone with internet access. However, these data archives mandate that data can only be stored there if it is biologically related. In the physical sciences, electron microscopy datasets can be submitted to general data repositories such as Zenodo, but repositories specific to electron microscopy are much easier to search as they can be archived with searchable metadata. EMPIAR has been funded by MRC and BBSRC grants, and similar funding from EPSRC for a physical sciences alternative would provide a world-leading resource for electron microscopy data. Our survey of UK research groups indicated that > 80% of those surveyed would be supportive of centralised data storage / archiving for published data.

A searchable data repository is not only vital to ensure EM data meets FAIR principles, but it would also act as a source of data for the development and training of new data analysis algorithms and workflows. One of the major developments of the last 8 years is the application of machine learning / artificial intelligence techniques to the analysis and simulation of EM data. The UK community has been at the forefront of the application of machine learning to electron microscopy, but the research is predominantly carried out in single research groups with that group's own data. In order to further develop deep learning analysis of data, there needs to be an accessible resource for sharing large experimental and simulated datasets that can be used for training machine learning algorithms. The data archiving of cryo-EM images has already had a large impact on the

development of data analysis workflows, where groups can develop and test new algorithms on datasets from world-leading facilities that they might not otherwise have access to.

In terms of the development of algorithms and workflows for data analysis, the UK EM community has been a leader in the development of open-source software for data analysis over the past two decades. This is exemplified in the primarily UK-based development of the Hyperspy Python package, which is world-leading open-source software for the analysis of data from scanning transmission electron microscopy (over 900 citations). However, over the last decade many of the key developers of Hyperspy have either left the UK or have left electron microscopy, putting the ongoing development of this critical open-source project at risk. Ongoing support for open-source data analysis software in physical sciences EM could be provided through the community applying for a collaborative computational project (CCP). These projects provide support for the development, maintenance and dissemination of software in a particular field. There is currently a CCP in cryo-EM (CCP-EM), which provides support for software development and training for single particle analysis. Creating a CCP for physical sciences EM would mitigate the risk of loss of support for the world-leading open-source projects within the UK and could act as a central hub for the development of the online data archive and machine learning advances described above. This idea was unanimously supported by respondents at the EMAG AGM in 2023.

The UK is a leader in the development of open-source data analysis for EM, but there are key aspects in which there is little development within the UK. One notable area that is lacking in UK development is EM image simulation. UK researchers typically use software (both proprietary and open-source) developed in European neighbours, the US and Australia to simulate electron microscopy images. The lack of simulation development in the UK is a risk to its world-leading position in terms of EM data analysis. Image simulation in particular underpins much of the analysis of atomic-resolution images, whether for image interpretation or in the testing of analysis workflows. This expertise should be further developed in the UK, which we suggest could be best developed through contributions to current open-source projects rather than the ground-up development of new simulation code.

A key issue for the future of EM data analysis in the UK is the training of electron microscopists. It was noted in the survey at the EMAG AGM that RTP staff are well-qualified to provide training on the practical use of microscopes but there is very mixed ability in terms of training in the use of software for data analysis (and very little in data simulation). There are training events in specific software at annual conferences, and workshops are run on an annual basis (e.g. annual workshop on the use of Hyperspy), but this provision isn't particularly helpful for day-to-day data analysis. It would be incredibly beneficial to researchers if current RTP staff at institutes were trained to provide training in data analysis and simulation, to provide in-house expertise at institutions where this is currently lacking. This train-the-trainers approach could be delivered through the proposed CCP or through the RMS/EMAG, where annual training courses would then lead to day-to-day support for researchers.

Recommendations

Data Management

- Community guidelines should be created for data storage for electron microscopy. This should include guidance for short-term and long-term options. The guidelines can be used by facilities (or individual groups) to cost data storage for funding applications.
- The physical sciences EM community should put in place a cloud storage platform for published data, similar to those in place in cryo-EM (EMDB and EMPIAR). This platform would provide a searchable database of EM data, which would promote the development of

novel data analysis algorithms/workflows (particularly machine learning algorithms that require training data).

- In order to enable published data on the searchable platform to be useable, published data should be stored in a common open-source data format. This format could be an already established open-source format such as EMD or Nexus, or could be a new community agreed format.

Data Analysis and Simulation

- The physical sciences EM community in the UK should apply for a collaborative computational project to ensure the maintenance of its world-leading open-source data analysis software. A CCP would mitigate the risk of key developers leaving the sector / UK and enable greater dissemination of the software to the community.
- Training in data analysis and simulation for RTP staff should be built into the calendar to ensure day-to-day support in these areas for researchers. A train-the-trainer approach would allow annual training courses to impact researchers at more institutions at the level of their individual research projects.

Note: Members of the physical sciences EM community applied for a new Collaborative Computational Project as part of the 2024 New Communities funding call but were unsuccessful in their application.

EM Roadmap: Data Analysis, Simulation and Storage Questionnaire Results

1. Does your facility provide support for users in terms of data processing?

Yes: 57% No: 43%

We have dedicated RTPs at offline workstations plus some method descriptions.

Some guidance by academics, group members and technicians associated with the data acquisition needing processing.

Partially but generally point users to packages and expertise and ask them to do it themselves, perhaps showing them once the procedure.

Graduate lecture courses and individual support.

We have an artificial intelligence research team of software engineers.

Training on standard software. Some provision of in-house developed code.

Basic analysis - instruction document shared with users so they can carry out some analysis. For more detailed analysis, we carry it out for them.

2. Does your facility provide support for users in terms of data management?

Yes: 43% No: 57%

Shared network drives for each microscope user to access.

Only data transfer but we store data for a limited period.
A network drive for users.

We keep a second copy of all raw data but tell the user they are responsible for their own data.

VMs and supercomputing, cloud storage.

This is left to research groups.

Archiving of open access data.

We backup all data but transfer data via a network drive and encourage users to make copies.

Time-limited storage on disk; support to transfer data to external storage.

A network drive for users.

3. Does your facility provide support for users in terms of simulation/modeling?

Yes: 19% No: 81%

We provide academic support but not real simulation / modelling support.

Not really but can advise on who to talk to.

Online courses / seminars. Analysis / modelling workshops. Online demonstrations.

Graduate lecture courses and individual support.

Some use of JEMS / multislice. Some collaboration for atomistic modelling.

Can discuss issues with users.

We don't do simulations / modelling.

4. How do you currently manage and fund data storage in your EM facility?

On a local server and microscope PCs.

University managed, funded by the facility.

University internal servers. Funded by faculty of sciences.

We take second copies not backups. They go on harddrives but the most recent data (3 months) go on a university network drive that is backed up.

Internal users have Imperial data storage and can either pay for more data storage or if linked with Diamond have further data storage capacity.

Store data on local server for limited period.

We don't. University has general protocols for data management and storage up to 1 Tb / group.

We store our data on the server managed by central University IT services.

Stored in University run HPC server.

Combination of local (NAS), cloud (O365) and cold (Azure) storage).

We have local storage, Dropbox for used data, also some LTS through University dMPs.

Managed by individual groups with funding from grants.

University controlled central data repository and local backups.

5. Are you supportive of a centralized service for data storage (for data linked to papers)?

Yes: 86% No: 14%

6. Would you be supportive of a funded network to support software development and training for EM analysis and simulation?

Yes: 100% No: 0%

7. Any further suggestions?

Methods for dealing with large data sets and machine learning etc would be useful.

A session at the facility managers' meeting (RMS) might be a good idea.

Network of expertise with possibly (part) funded positions (and training) of permanent facility staff.

Larger repository locations. Better agreement on required metadata.

Better networking to share needs and solutions.

Recordings of hyperspy school would be useful to be able to refer back to.

There are many open source resources available, but signposting is patchy. Indexing of these would be helpful.

Standard metadata. Maintain strong open-access behaviours.

Improve data sharing for experiment sharing.

Try to collaborating with development in other disciplines.

Feel there would be a benefit to tiered workshops for python-based analysis. Workshops at conferences etc. now seem to be covering a very broad range of users from beginners to experience users of eg. Hyperspy. Is there mileage in having separate workshops for beginners, intermediates etc, or ones dedicated to e.g. 4DSTEM, spectroscopy etc?

Centralised storage for published data.