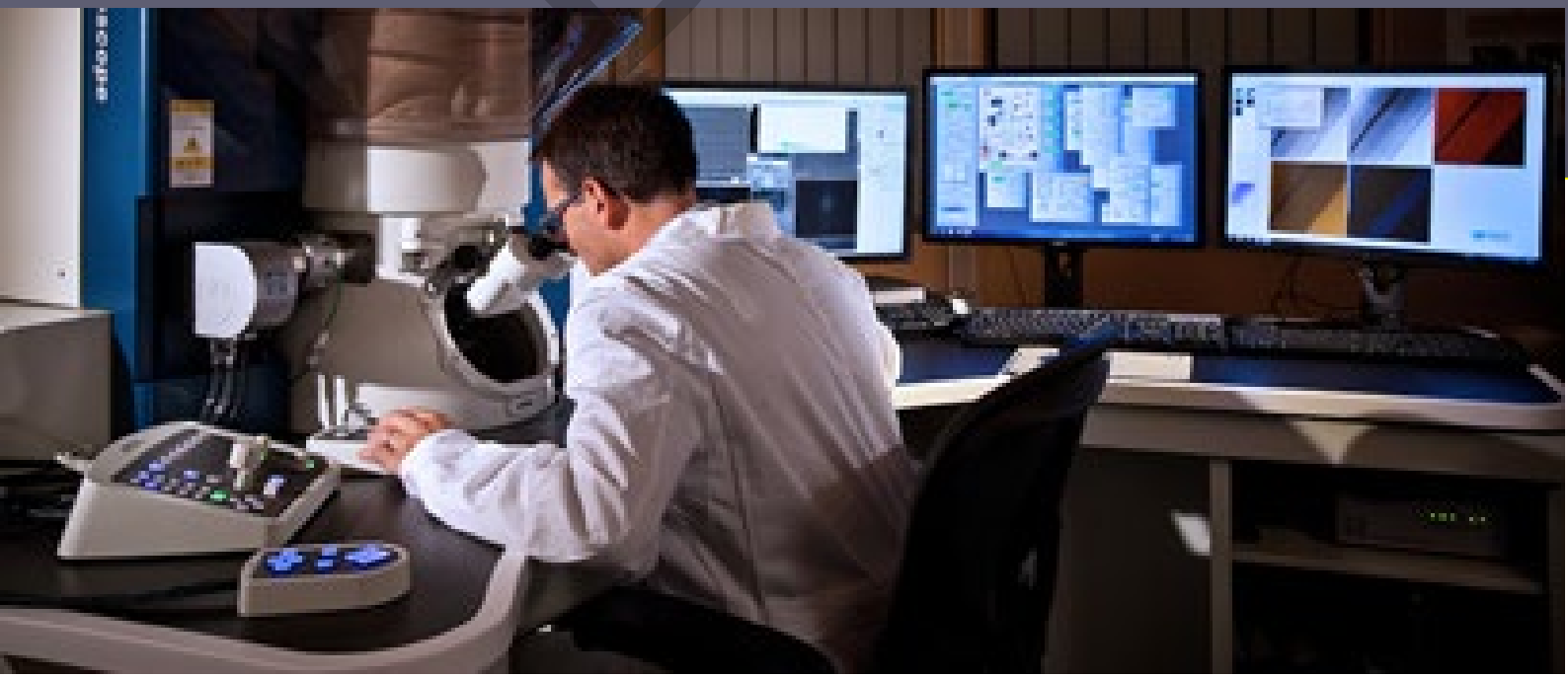
esteem3

Enabling Science through European Electron Microscopy

Final report on a data and software strategy for TEM

Deliverable D1.8 – version 1.1

Estimated delivery date:	31st April 2023
Actual delivery date:	11th April 2023
Lead beneficiary:	JUL
Person responsible:	Dieter Weber
Deliverable type:	☒ R ☐ DEM ☐ DEC ☐ OTHER ☐ ETHICS ☐ ORDP
Dissemination level:	☒ PU ☐ CO ☐ EU-RES ☐ EU-CON ☐ EU-SEC



THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON 2020 RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO **823717**

Grant Agreement No:	823717
Funding Instrument:	Research and Innovation Actions (RIA)
Funded under:	H2020-INFRAIA-2018-1: Integrating Activities for Advanced Communities
Starting date:	01.01.2019
Duration:	54 months

Table of contents

Revision history log

Version number	Date of release	Author	Summary of changes
V0.1	2023-03-31	Dieter Weber	Initial version for review
V1.0	2023-03-31	Peter A. van Aken	Minor changes and approval
V1.1	2023-04-11	Aude Garsès	General review

Reference to previous version

This report is an update and extension of the preliminary report (deliverable 1.4a). Only updates are included here.

Description of Task 1.4

Quoted from the proposal:

> *The development of a strategy for open data and open science within the ESTEEM3 infrastructure in a big data context that is compatible with current European directives, including the selection of data and software that will be made open (raw data, metadata, processed data, complete/partial data from experiments, etc.), embargo period, practical implementation and sustainability (duration)*
>
> *The development of a data strategy for user (TA) access to the ESTEEM3 infrastructure in terms of data retrieval, compatible file formats and data standards, open software platforms, post-processing facilities and approved approaches for archival data storage, including the development of tools to convert data between different file formats*
>
> *An evaluation of open software for TEM, including a survey of available platforms (user base, license, maturity), verification of compatibility with EU open license (EUPL), evaluation of added value for users, evaluation of sustainability (cost, maintenance, user support and durability) and a study of hybrid (open/proprietary) solutions.*

Strategy for open data and open science

The positive developments described in the preliminary report continue. The following case study for event-based detectors can serve as an example.

Event-based detectors for 4D STEM

ESTEEM3 partner ANT has shown that event-based detectors that register the impact of individual electrons can allow very high scan speeds in 4D STEM[1]. ESTEEM3 partner JUL is contributing to the LiberTEM[2] open source project for high-throughput (S)TEM data processing, also supported through ESTEEM3 WP 11. Amsterdam Scientific Instruments (ASI) has developed a Timepix3-based detector for TEM.

After the work of ANT showed the potential of event-based detectors for 4D STEM, ASI and JUL collaborated to develop practical application support for this and test it for different scientific use cases. Event data with time and coordinate of impact of individual electrons is rather different in structure compared to the usual frame-based data of conventional image detectors. Most software for analysis of conventional 4D STEM data represents the data as an n-dimensional numerical array. Sparse arrays, and in particular the popular Compressed Sparse Row (CSR) format, were identified as a suitable link between event-based data and numerical arrays: Only non-zero values and their positions, corresponding to electron impact events, are stored in this data format, but to the user the interface of an n-dimensional numerical array is presented, corresponding to a 4D STEM dataset. Sparse arrays are well-established in scientific computing, and software libraries are available for such

[1] https://doi.org/10.1016/j.ultramic.2021.113423
[2] https://libertem.github.io/

data that expose an API that is compatible with dense arrays. That means in may cases sparse arrays can be used as a drop-in replacement for dense arrays in data analysis codes, including 4D STEM applications.

Within the collaboration, ASI implemented conversion of event data to the CSR format, and JUL implemented support in LiberTEM to read and process such data. Through the previously developed integration between HyperSpy and LiberTEM (see reports from WP 11), it is also possible to process the data with pyXem, another software package for 4D STEM, where ESTEEM3 partner TRO is a major contributor.

Through these developments, it is now possible to process data from ASI's detector just like any other 4D STEM data, just with much smaller files, faster scan speed, and faster processing for routines with native sparse array support.

This case study shows how engagement of research institutes in open source software and active collaboration with vendors, along the mission of WP 11 in ESTEEM3, can have a positive impact on open science and open data, and at the same time support development of innovative products by small and medium enterprises (SMEs).

Open source solutions for TEM data analysis like HyperSpy, pyXem and LiberTEM have become mature and widespread enough that detector vendors can lean on them for scientific application of their products, meaning they don't have to provide a full application software stack if they open their interfaces and file formats for integration with such open source solutions. Only through their long-term involvement in such projects did JUL have the necessary know-how to participate successfully in such a collaboration

Since ASI and JUL collaborated early-on, they could define a file format and detector interface that is compatible with open science and open data principles, and at the same time works well in practice. Previously, file formats and detector interfaces were defined mainly by detector vendors. Since the chosen format is based on the well-established CSR format that was developed in the 1960s, existing software libraries make it comparatively easy to integrate support for it in different software and ensure that the files can likely be interpreted also far in the future.

Since the contribution of JUL to this collaboration is available under an open source license, third parties can use it under fair, reasonable, and non-discriminatory (FRAND) terms. That allows other detector vendors to also use the same format and software support. It also makes multilateral collaboration very easy since the license terms are clear and well-established, and give all involved parties freedom to use the products of the collaboration for their purposes.

Key to this project are established standards: The CSR format and the NumPy ndarray API that allows using sparse arrays in the place of dense arrays in Python-based applications with only limited code changes.

In conclusion, this project shows how important long-term support and involvement of research facilities in software and application development are. Public funding for such activities, like ESTEEM3 WP 11, is a critical enabler for this and actively contributes to the goals of ESTEEM3, namely open science and open data.

File formats for TEM

Open source file reader support for raw data continues to improve. By now most formats can be opened with a range of different applications, and conversion is possible as well. The RosettaSciIO[3] project separated the HyperSpy file readers from application code to make them re-usable and share them between different projects.

For technical metadata such as instrument parameters the situation remains difficult. Several different activities are attempting to formalize and standardize metadata for TEM data, but for now there doesn't seem to be traction for practical application. TEM metadata is on the difficult side compared with other modalities, since the instrument is so flexible and at the same time unstable. It has a high number of control channels, a complex transfer function (relation between specimen properties and measured data), and only a loose connection between external control input to the instrument and the instrument's transfer function. Drift requires frequent recalibration, and it means the instrument's transfer function may even change during acquisition of a dataset and has relatively high uncertainties.

At ESTEEM3 partners and their collaborators, there's ongoing work for digital twins that can replicate the transfer function of a given instrument, for example TEM Gym Basic[4], methods to calibrate the transfer function of a real instrument, and ensuring consistent and well-defined behaviour of software for data analysis. See also WP 11 Task 11.4.

Open software and data

Publication of relevant software and data together with a scientific paper is becoming more and more widespread in electron microscopy. It could be time to start a discussion if one should formally require this for publication in relevant journals: As computational methods continue to grow in importance for electron microscopy, a paper can often only be meaningfully reviewed, reproduced and used as a basis for follow-up work if software and data are available. Furthermore, software bugs and/or incorrect use of software can make scientific results invalid or incorrect, meaning the used software should be reviewable, in particular if it is newly developed.

That also means that good software development practices (issue tracking, peer review of code changes, unit tests, change management, quality assurance) become important. The review criteria of the Journal of Open Source Software[5] can be an example for minimum requirements for research software engineering.

Furthermore, the collection of TEM and STEM data in open repositories that allow data science studies and comparisons is growing. A recent paper on a ptychography method improvement at JUL[6] can serve as an example for a paper that relies heavily on previously published code and data, and in turn makes the resulting code available. Here, selected high-quality datasets that are linked to a peer-reviewed publication and include a detailed description and references to parameters and software are most valuable. With the TEM metadata issues described in the previous subsection, it is still not sensible to publish data *en masse*, since detailed and universally understood metadata formats to interpret such a "data dump" are still not available.

[3] https://hyperspy.org/rosettasciio/
[4] https://temgymbasic.readthedocs.io/en/latest/
[5] https://joss.readthedocs.io/en/latest/review_criteria.html
[6] https://doi.org/10.1093/micmic/ozad021

Data strategy for TA users

In the previous report, requirements for a data management system were lined out. A system that can fulfil these requirements is now operational at JUL[7]. It is modular and extensible based on open source components and standardized protocols to connect them. Facilities for data analysis, such as servers that can run Jupyter notebooks, workstations with software such as Digital Micrograph, and the web GUI of LiberTEM, are directly connected to the data management system at JUL. If desired, TA users can obtain access to these facilities, including remote access from their home institution. Alternatively, they can access files via sync and share through Nextcloud or receive a copy on external storage media. As a next step, integration of electronic lab notebooks with the data management is underway at JUL. Currently the system is customized for application at JUL and might require changes or extensions to also work at other facilities, but it is planned to make it easier to deploy at different sites for wider application. JUL is happy to share it on request.

With open source or free-of-charge data analysis solutions, TA users can install the required software stack on their systems. Help with the software setup and an introduction to the tools can be part of TA. Since software like HyperSpy, pyXem and LiberTEM can process large datasets efficiently out-of-core, TA users only require a normal PC to analyse also large datasets.

Evaluation of open software for TEM

The list of software from the previous report is still largely up-to-date. Notable changes:

- Pixstem merged with pyXem.
- RosettaSciIO[3] is a new HyperSpy spin-off for the file reading code to make it easier to reuse.
- Py4DSTEM[8] should be added to the list. It is licensed under GPL-3.0.

Conclusion and outlook

The positive developments for software and data in electron microscopy in an open science and open data context continue.

Metadata for TEM remains a difficult topic, however. Here, the main challenge will be to achieve practical usability, wide acceptance, compatibility, and unambiguous interpretation across a very heterogeneous hardware and software landscape.

Both funding and appreciation of research software engineering by scientists and staff at research facilities with activities like ESTEEM3 WP 11 will be critical for continued progress. These activities help to bridge the gap between “works in principle” and “works in practice”. Know-how in this area is also a key success factor for participation in (open) innovation in collaboration with commercial players. Furthermore, open source solutions created with support from public funding incentivize openness and interoperability in a software ecosystem, as highlighted in the case study on event-based detectors, as opposed to incentives for “walled gardens” if vendors have to shoulder the entire burden of development for a product and finance it with sales.

[7] https://er-c-data.fz-juelich.de/
[8] https://py4dstem.readthedocs.io/en/latest/index.html