Ref. Ares(2019)7022347 - 11/12/2019

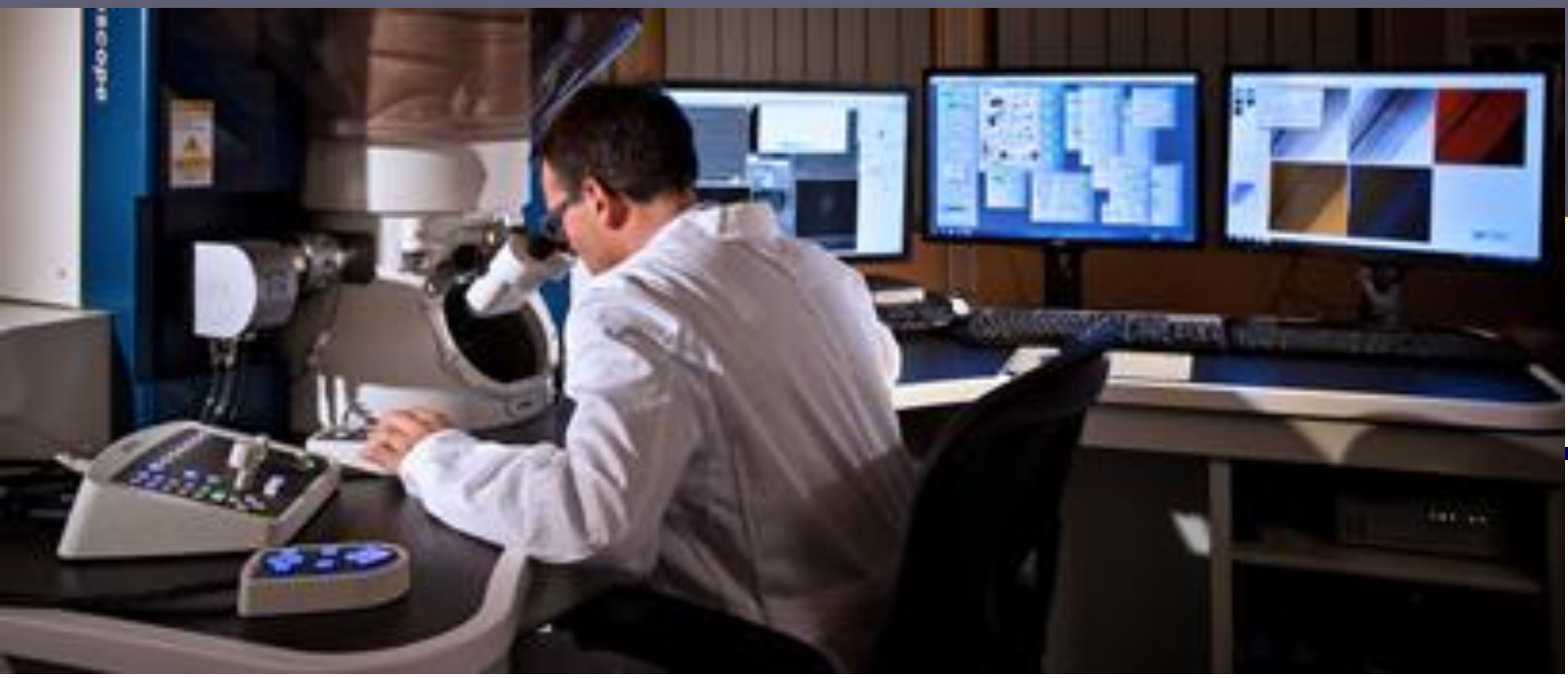
esteem3

Enabling Science through European Electron Microscopy

Initial report on a data and software strategy for TEM

Deliverable D1.7 – version 1.3

Estimated delivery date:	31/12/2019
Actual delivery date:	11/12/2019
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:	48 months

Table of contents

Revision history log

Version number	Date of release	Author	Summary of changes
V0.1	2019-10-15	Dieter Weber	Initial version
V0.2	2019-10-22	Dieter Weber	Feedback from Vadim Migunov
V0.3	2019-11-08	Dieter Weber	Feedback from Antonius T. J. van Helvoort and Duncan Johnstone
V0.4	2019-11-13	Dieter Weber	Feedback from Rafal E. Dunin-Borkowski
V1.0	2019-11-13	Dieter Weber	Final version
V1.2	2019-11-18	Lucie Guilloteau	Editing
V1.3	2019-12-11	Peter van Aken	Final approval

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."*

Summary of requirements for scientific TEM software and data

- Data acquisition:
  - Acquisition software should make data available in suitable forms to allow versatile processing, including FRAND access to the data under full control of the user.
  - Well-documented data format with an appropriate choice of readers that are available under FRAND terms, ideally Open Source under a permissive license to allow re-use of the code and integration in a wide range of software packages.
  - Access to raw detector readout. Simple standard post-processing such as dark frame subtraction and gain correction should be optional features and must be verifiable. Advanced processing such as denoising and distortion correction must be optional, preferably as part of data processing and not data acquisition.
  - In the future, access to live data streams through suitable open APIs, since live data processing will continue to grow in importance.
  - In the future, standardized metadata to allow automated processing, indexing and management of repositories.
    - For the time being, before a standard has been defined and accepted, well-documented metadata is sufficient.
- Data processing:
  - Verifiable data analysis routines to ensure that analyses are correct and that "black box" solutions are not used.
  - Extensibility to enable innovative work.
  - Accessibility under FRAND terms to ensure that third parties can replicate analyses.
- Publication:
  - Combination of input data with precise descriptions of analysis workflows leading to published results, where ideally:

        - the input data are raw detector data
        - descriptions are executable scripts or programs that perform transformations from input data to results in a verifiable, unambiguous and clear manner, together with human-readable descriptions.
- Long-term sustainability:
    - Data must remain interpretable for the long term for archiving purposes, including the necessary software tools.
    - Software tools that are necessary to operate instruments and to perform innovative analysis require frequent maintenance and improvement, such as adapting to changes in underlying platforms and new hardware or processing routines.

Existing solutions

The following table provides an non-exhaustive overview of current software packages for electron microscopy. The column “ESTEEM3 Partners contributing” refers to individual scientists who contribute to these software packages and are affiliated with ESTEEM3 partners, but not to an official relationship between the partner institution and the software development project or to funding of the software through ESTEEM3.

Name	Description	URL	License	ESTEEM3 Partners contribu-ting	User base	Maturity
Gatan Microscopy Suite	Gatan Microscopy Suite (GMS) sees itself as industry standard software for (scanning) transmission electron microscopy experimental control and analysis.	http://www.gatan.com/products/tem-analysis/gatan-microscopy-suite-software	Proprietary	GRA TOU (plug-ins)	Very large	Very mature
HREM Research	A dedicated site selling software for quantitative electron microscopy, often in the form of plug-ins for the Gatan Microscopy Suite.	https://www.hremresearch.com/index.html	Mixed	TOU (GMS plug-ins)	-	-
Hyperspy	An open source Python library, which provides tools to facilitate interactive data analysis of datasets that can be described as multi-dimensional arrays of given signals (e.g., 2D arrays of spectra, a.k.a. spectrum images).	https://hyperspy.org/	GPL v.3	CAM OXF ANT JUL TRD	Large	Mature
pyXem	An open-source Python library for crystallographic electron microscopy. The code is developed primarily as a platform for hybrid diffraction-imaging microscopy based on scanning electron diffraction (SED) data.	https://pyxem.github.io/pyxem-website/	GPL v.3	CAM TRD	Medium	Stable
Pixstem	A library for processing data acquired on fast pixelated electron detectors using scanning transmission electron microscopy (STEM).	https://pixstem.org/	GPL v.3	ANT	-	-
LiberTEM	An open source platform for high-throughput distributed processing of large-scale binary datasets using a simplified MapReduce programming model.	https://libertem.github.io/LiberTEM/index.html	GPL v.3, parts MIT	JUL	Small	Beta

ASTAR	An automatic crystallographic indexing and orientation/phase mapping tool for TEM.	https://www.nanomegas.com/Documents/ASTAR.pdf	Proprietary	NM	Large	Mature
Nion Swift	Open source scientific image processing software, integrating hardware control, data acquisition, visualization, processing and analysis using Python.	https://nionswift.readthedocs.io/en/stable/	GPL v.3		-	-
Thermo Fisher Velox	Acquisition and analysis software with hardly any information publicly available.		Proprietary		Small	-
Other Thermo Fisher software	Hardly any information publicly available.		Proprietary	JUL	Large	Mature
Amiro-Avizo	3D visualization and analysis solution for scientific and industrial data.	https://www.thermofisher.com/de/de/home/industrial/electron-microscopy/electron-microscopy-instruments-workflow-solutions/3d-visualization-analysis-software.html	Proprietary		-	-
JEOL software	Hardly any information publicly available.		Proprietary		Large	Mature
Hitachi software	Hardly any information publicly available.		Proprietary		Medium	Mature
SerialEM	Software that can be used to acquire a variety of data from electron microscopes, including tilt series for electron tomography, large image areas for 3D reconstruction from serial sections and images for the reconstruction of macromolecules.	http://bio3d.colorado.edu/SerialEM/	MIT		Large	Mature

Instamatic and Problematic	Instamatic is a Python program that is being developed with the aim to automate the collection of electron diffraction data. Problematic is a collection of routines for processing serial electron diffraction data collected using Instamatic.	https://github.com/stefsmeets/instamatic https://github.com/stefsmeets/problematic	GPL v3/2		-	-
Pycroscopy	Software for the scientific analysis of imaging data.	https://pycroscopy.github.io/pycroscopy/	MIT		-	-
Prismatic	CPU / GPU software for the fast simulation of scanning transmission electron microscopy (STEM) experiments.	https://prism-em.com/	GPL v.3		-	-
MULTEM	A collection of routines written in C++ with CUDA to perform accurate and fast multislice simulations for different TEM experiments.	https://github.com/Ivanlh20/MULTEM	GPL v.3		-	-
Dr. Probe	A tool package for STEM and TEM image simulation.	https://er-c.org/barthel/drprobe/	Mixed proprietary and GPL v.3	JUL	-	-
EMsoft	A series of programs along with a library, mostly written in Fortran 90, with some OpenCL contributions, for the computation and visualization of scanning electron microscopy diffraction patterns.	https://github.com/EMsoft-org/EMsoft	3-clause BSD	TRD	-	-
ImageJ and Fiji	General purpose image processing software with a large number of microscopy-related data analysis plug-ins, including tools for image filtering, video composition, size distribution analysis and tomography.	https://imagej.net/Fiji	Mixed 2-clause BSD and GPL		Very large	Mature
MTEX	A free Matlab toolbox for analyzing and modeling crystallographic textures using EBSD or pole figure data.	https://mtex-toolbox.github.io/	GPL v.2	TRD	-	-
MacTampasX, CrystalKitX	TEM and STEM data analysis software	https://www.totalresolution.com/	Proprietary		-	-

Compatibility with EUPL

This summary does not provide legal advice and is included for discussion purposes.

EUPL is a copyleft license, which is approved by the Open Source Initiative (OSI), falling under their definition as Open Source. It states to be compatible with several other copyleft licenses that are listed in its appendix, including GPL v.2 and v.3. EUPL-licensed code can be included in works under those other copyleft licenses. Code under more permissive licenses, such as Apache or MIT, can also be included in works under such copyleft licenses. Copyrighted material under copyleft licenses cannot be included in works that are licensed under more restrictive terms, including proprietary software and software that is available under other incompatible terms, such as free for academic use only, as in the case of CCP4 http://www.ccp4.ac.uk/ccp4license.php. In contrast to copyleft licenses, software that is released under more permissive licenses, such as MIT or Apache, *can* be included in works that are under more restrictive terms.

From a scientific user's perspective

In the context of the requirements listed above, code that is relevant for ensuring compatibility or interfacing between different pieces of software should be licensed under permissive terms and not copylefted, in order to enable widespread use. In the context of electron microscopy, this includes support for file formats and interfaces. This licensing model allows exploitation in proprietary software without dual licensing.

For scientific code that should be protected against unreviewable modification or exploitation in proprietary software, copyleft licenses are appropriate and fulfil all of the requirements listed above.

Patents have been a concern for open source software, since they can undermine licensing that is based on copyright. It is now common practice to use open source licenses that include a patent license.

Licensing under more restrictive terms should be chosen with great care, since it can make FRAND access and verification more difficult. Licensing under fair and compatible terms is essential for collaboration in the community and for dissemination.

Impact on market for electron microscopy solutions

Open Source solutions have the potential to disrupt a market that is based on commercial licensing, since their business model can become very efficient through their economy of scale once a critical mass of users and developers is reached. As an example, GNU/Linux made commercial licenses for UNIX systems nearly obsolete within a few years.

In electron microscopy, companies play a critical role to develop, produce and maintain microscopes, accessories such as detectors and sample holders, as well as the software that is required to operate them and interpret data. Furthermore, they help to make innovations from academia available to a wide audience. Ensuring a smooth transition and long-term sustainability of commercial products and services, which are an important asset for electron microscopy, in an environment that includes more Open Source solutions, is an important task for the electron microscopy community. Licensing and business models should foster a fair partnership between companies and users with a rich ecosystem of available solutions and a functioning market.

Sustainability models for Open Source software

The maintenance of Open Source software to ensure long-term sustainability requires an appropriate business model to fund development work. Broad redistribution and modification rights for Open Source software, which are ideal from a scientific perspective, are at odds with traditional commercial licensing models. Alternative models are therefore required.

Dual licensing with a copyleft license and a commercial license is in principle possible. However, it makes collaborative development difficult, while still giving rather unrestricted access to the product free of charge. It combines the disadvantages of both licensing models, without retaining their advantages.

The following strategies can help to sustain software development and to foster a fair and open collaboration between companies and scientific users:

- A business model in which selling hardware is not affected. Some companies see software as a necessary enabler for hardware sales, i.e., a cost centre and not a profit centre. The use of Open Source solutions allows companies to reduce their software development costs. The opening and documentation of interfaces so that Open Source software can interface with their solutions can be of mutual benefit.
- Moving from traditional software licensing to providing services such as installation, maintenance, customization and hosting.
- Mutual alignment of architecture and licensing models between scientific and commercial users.
    - Licensing under permissive Open Source terms such as BSD, Apache or MIT licenses if components should be included in proprietary software, for example to support standards.
    - Suitable architectures to create interfaces between copylefted and proprietary software that are compatible with respective license terms. Possibilities include plug-ins, scripting languages, open APIs and network services, allowing proprietary software to be connected to software that is licensed under more restrictive terms such as GPL.
    - Use only licenses that allow commercial use, i.e., OSI-approved Open Source licenses.
- The mission of some large organizations, such as the Helmholtz Association in Germany, includes supporting large-scale scientific infrastructure for the long term, including software, creating an avenue for the long-term sustainability of scientific Open Source software.
- EU funding (e.g., within ESTEEM3) supports Open Source software development, including as a form of commercialization.
- Support for workshops sponsored by hosting organizations, for example alongside conferences.
- Several foundations and large companies support Open Source software projects that are important for their mission or business, most commonly within the IT industry.

In summary, widely- used and well-understood OSI-approved open source licenses with or without copyleft that include a patent license are ideal for collaboration and dissemination in the context of scientific software and data. The development of a model for the long-term sustainability of Open Source software, as well as ensuring a functioning ecosystem between Open Source solutions and commercial options, is an ongoing process.

Selection of data and software that will be made open

Based on an initial screening, Zenodo is currently regarded as the best suited data sharing platform for the electron microscopy community. It makes data FAIR and is flexible with regard to data type. Several TEM data sets have already been deposited there, including work that has been marked as related to ESTEEM3: https://zenodo.org/search?q=ESTEEM3

Electron microscopy data sets and their interpretation are closely tied to specific samples, preparation and imaging conditions. Furthermore, experiments can be error-prone, generating unusable data. For these reasons, it makes sense to only publish selected data sets, instead of releasing all data indiscriminately. Such a selective strategy has already been followed successfully on Zenodo.

The establishment of data repositories for electron microscopy data, just as for other fields such as genome sequencing in life science and X-ray powder diffraction, is an ongoing process. Since data acquisition and processing in electron microscopy allow for many variations, reliable automated data interpretation has proven to be more difficult than in other fields and is currently an active and dynamic research field.

The publishing of raw data together with software, for example in the form of a Jupyter notebook that can be used to convert raw data into final, processed results in a transparent and reproducible manner, is proving to be a viable strategy for electron microscopy data.

Since this is a topic that is developing dynamically, ESTEEM3 partners are encouraged to experiment with various models and to share their experiences, in order to develop best practices over time.

Data strategy for user (TA) access to the ESTEEM3 infrastructure

Since novel methods in electron microscopy can generate very large data sets, it can be advantageous to store them at the facilities where they were generated, instead of transferring them. The availability of data processing facilities alongside electron microscopes for TA users is likely to be highly beneficial in the future, as faster detectors produce larger data sets, which require specialized high performance hardware and software for practical processing that may not be readily available to all users.

Scientific communities that work on high energy physics and at other large-scale facilities have needed to address issues with data access and processing for many users. However, the cultures and traditions of who manages and owns data can be different in each community. For example, data management in high energy physics is traditionally the responsibility of the facility and of beamline and instrument officers, who perform or supervise experiments, provide copies of the data to users and give them access to large-scale data sets. They may help with processing data recorded from specialized instruments. In electron microscopy for materials science users still often bring their own storage media and are personally responsible for the appropriate management of their own data, with many different users working on the same electron microscopes. In contrast, in cryo electron microscopy in the life sciences there is greater standardization and uniformity in data handling, which will be described in future updates of this document.

Next-generation data handling for electron microscopy has to bridge these two worlds. On the one hand, data should stay at facilities and be managed in a centralized fashion, in order to ensure long-term archiving and backup, alongside offering the infrastructure to handle and process large data sets that can no longer be handled easily on external storage media and personal computers. On the other hand, users should be able to manage their own data, including the provision of access to collaborators and publishing datasets on appropriate platforms. These changing circumstances mean that facilities will have to plan for long-term archival storage, including making provision to cover such costs. Support from research funding agencies is required alongside viable technological and organizational solutions, including the possibility to delegate long-term archival storage and backup to third parties. Furthermore, facilities must work towards complying with various requirements for data management from the different agencies that fund individual projects. Ideally, such data management requirements should be consolidated internationally, in order to facilitate compliance for both individual users and facilities.

The implementation of such policies in practice is an ongoing process at the facilities of the ESTEEM3 partners. Aspects that require further clarification include financing, the assignment of responsibilities within each institution and with third parties, requirements analysis, as well as the choice and availability of suitable hardware and software. In Australia, such a workflow is deployed: https://static1.squarespace.com/static/5594e714e4b0f2448bc41c42/t/5bdaccda352f5342623c45ad/1541065979119/MyTARDIS-Monash-Wojtek.pdf

A practical system that addresses users' needs should allow personal-computer-based processing of smaller files, alongside centralized high performance processing of larger files and computationally intensive applications.

Other aspects of the present situation are as follows:

- The choice and definition of standards for file formats are a work in progress that is difficult to resolve, since electron microscopes are highly configurable and dynamic, making automated quantitative data interpretation difficult. Work package 11 in ESTEEM3 addresses aspects of this problem.
- Open software platforms are available and are under constant development, including at the facilities of the ESTEEM3 partners. The use of compatible licenses, such as GPL 3.0, ensures that many of these packages can interoperate with each other, as well as share or exchange code. They are sustained by a range of funding sources, including ESTEEM3 itself. Work package 11 in ESTEEM3 covers targeted developments in this area.
- Open Source readers for many file formats are available. Major vendors have started to welcome and support their development, for example by providing format specifications and providing permission to publish them under Open Source licenses.
- Tools such as Hyperspy can be used to read and write different file formats, allowing conversion between them.

Ongoing activities within ESTEEM3

- Several Open Source solutions are under development by members of the ESTEEM3 consortium, in particular as part of work package 11. Examples of such developments include:
    - Pixstem https://pixstem.org/
        - A library for processing data acquired using fast pixelated electron detectors.
    - pyXem https://pyxem.github.io/pyxem-website/

        - Application-driven crystallographic analysis of 4D-SED/4D-STEM data, including crystal phase and orientation mapping, strain mapping and non-crystalline phase mapping based on pair distribution function analysis.
    - LiberTEM https://libertem.github.io/LiberTEM
        - APIs and processing back-ends for interactive live processing at high data rates with low latency on distributed systems.
    - Active work on interfacing with other solutions, including commercial products.
    - Activities focusing on interoperability and avoiding duplication:
        - Several ESTEEM3 partners are involved in software projects for electron microscopy for similar applications, such as 4D STEM.
        - Interoperability is a common goal, with frequent contact between groups.
        - The use of different own platforms in different groups ensures freedom to operate and to try new approaches, since each partner has different interests and ideas.
        - Open Source licensing allows the re-use and exchange of components.
    - In general, there is a healthy ecosystem that is under constant developments
- Ongoing work on practical solutions to facilitate next-generation electron microscopy data workflows, including:
    - A requirements analysis has been completed at the JUL partner, which is currently evaluating and prototyping solutions.
    - Advanced systems, such as the Australian cloud, cannot be taken over directly because of different boundary conditions, including a different scale compared to the European Union. However, they can serve as an inspiration and parts of them can perhaps be re-used. Differences in the European Union include the fact that:
        - Australia has a more centralized approach to managing microscopy centers through the Microscopy Australia organization, whereas the European landscape is more heterogeneous.
        - There is a much smaller user base in Australia, which makes establishing centralized solutions easier. In Europe, a much larger number of players has to be coordinated.
    - Other facilities should be described in this document in the future.
- Several ESTEEM3 partners, including ANT, GRA, TRD and JUL, are developing best practices to make data and software follow FAIR principles, including:
    - Uploading data to Zenodo, sometimes in combination with Jupyter notebooks to document processing workflows
    - https://zenodo.org/record/2563880
    - https://zenodo.org/record/2566137
    - https://zenodo.org/record/2578866
    - Since files can be very large, efficient storage (e.g., through compression), as well as fast write and read access to data, are important, further highlighting the need for improved data formats.
- Metadata for electron microscopy is a challenging topic, since electron microscopes are complex, versatile and dynamic instruments to a much greater extent than many other scientific analysis tools. Relevant activities in this direction include:
    - A requirements analysis: https://github.com/LiberTEM/nexus-4dstem/wiki/Requirements-analysis
    - The possibility of using the NeXus format https://www.nexusformat.org/ as a basis.
    - Ongoing links to the FAIRmat initiative: https://fairdi.eu/fairmat/consortium

Summary and outlook

In conclusion, there are positive developments in data and software strategies for electron microscopy, including increasing support for paradigm changes towards open, interoperable solutions. Several Open Source projects promise to establish themselves and to achieve long-term sustainability, interfacing both with each other and with proprietary solutions. Clear political directives towards Open Science and Open Data, together with the support and dedication of funding agencies and of individual contributors, will be instrumental for working towards these goals.

The activities described in this document will be developed and refined further as the ESTEEM3 project progresses. A final report will be written as Deliverable 1.7b.