LUND UNIVERSITY

Research Data Management from planning to analysis: A case study of the NanoMAX beamline station at MAX IV Laboratory

Johnsson, Maria; Lassi, Monica

2018

Document Version:
Publisher's PDF, also known as Version of record

Link to publication

Citation for published version (APA):
Johnsson, M., & Lassi, M. (2018). *Research Data Management from planning to analysis: A case study of the NanoMAX beamline station at MAX IV Laboratory*.

Total number of authors:
2

General rights
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: https://creativecommons.org/licenses/

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Research Data Management from planning to analysis: A case study of the NanoMAX beamline station at MAX IV Laboratory

Maria Johnsson, maria.johnsson@ub.lu.se and Monica Lassi, monica.lassi@ub.lu.se
e@LU – eScience and eInfrastructure, Lund University
August 2018

Introduction

This report summarises a small case study in which two research groups were interviewed and observed as they were planning, conducting and analysing experiments at MAX IV Laboratory[1], a Swedish national laboratory for synchrotron X-rays available for international users. The aim of the study was to identify the current research data management (RDM) practices of the scientists, and to connect this to how these practices are supported by MAX IV at present. RDM entails the management of research data throughout its life cycle; from planning a project, through to collecting, processing, analysing, storing, archiving, and sharing research data. In this study, we were interested in how the current tools for documentation (including metadata), and data storage together support the scientists' research processes. The set of tools and support for managing research data in different stages of the research data life cycle are often denoted as research data services.

The primary audience for this report is MAX IV, in particular the Controls & IT department, the NanoMAX beamline, and the User Office who were engaged in the planning of the study and during our visits to the facility. Other audiences that might find the report useful include people who are developing research data services, or are interested in gaining a basic understanding of how scientists visiting a synchrotron facility work with research data. Therefore, the report includes some contextual information about the organisation and some synchrotron specific nomenclature that might need explaining. The study was conducted within the e@LU - eScience and eInfrastructure project at Lund University, which included a sub-project on developing services for metadata, ontologies and data provenance. e@LU was funded by Lund University and initiated by eSSENCE[2], a strategic collaborative research programme in e-science with Lund University, Umeå University, and Uppsala University as collaborators.

The results of the case study can inform the development of research data services for MAX IV, focussing on the practices and experiences of users. Developing research data services can be seen as an iterative process, in which the users participate to test and give feedback; thus this study can be thought of as an early stage in such an iterative process. Suggestions for building on this work is provided in the final part of this report. The following sections of the report are: Study design, Findings, and Conclusions including suggestions for future studies. The Findings section contain suggestions for further development of the research data services, based on the users' accounts of their visits at MAX IV.

---

[1] MAX IV Laboratory, https://www.maxiv.lu.se/
[2] eSSENCE, http://essenceofescience.se/

Study design

The aim of the study was to identify the current research data management (RDM) practices of the scientists, in relation to how these practices are supported by MAX IV at present. The question that guided the study was: “How do scientists work with research data and metadata in the planning stage, during the experiments, and in the early phases of data processing and data analysis?”

The case study was designed in collaboration with the Controls & IT department, the User Office and the beamline scientists at the NanoMAX beamline station. NanoMAX is up and running, but is still under development. This means that the users visiting the station provide important input into its continuous development. Further, NanoMAX has a diverse user community, in that users are active in many different research areas, experiments are guided by different types of research questions, and that there is not a standardised way of managing research data within the user community. Some of the planned and running beamline stations at MAX IV have more homogenous user communities with well-established RDM practices.

Study participants were recruited with the help of MAX IV staff, who filtered out the visitors who had been awarded beamtime at NanoMAX, i.e. been allocated a time slot for conducting experiments, and who were also employed at Lund University. The study participant selection was done to allow for synergy effects, as Lund University constitutes a strong user base for MAX IV, and is in the process of developing university-wide research data services. Data were collected through semi-structured interviews and observations. The study participants gave their informed consent on a form that, along with other relevant material, will be archived in the University Archive.

To respect their integrity, we henceforth call the two cases studied “Case1” and “Case2”. The two cases are situated in different research areas, with Case1 coming to the beamline with a biological research question, and Case2 researching methods in materials science. In both cases, the applicants of the beamtime were the Principal Investigators (PIs), i.e. the leaders of the research groups.

Before the beamtime, semi-structured interviews were conducted with each case’s PI. The first interview focussed on the study participants’ ways of working with research data in general when conducting experiments at synchrotron facilities; how they were preparing their samples; and their expectations for the experiments at NanoMAX.

During the beamtime, we conducted observations using a semi-structured observation protocol, focussing on how the study participants were interacting with the IT tools at the beamline to document their experiments.

After the beamtime, one to two months afterwards, we followed up with a second semi-structured interview with the applicant of the beamtime. In one of the cases, we conducted an extra interview with a postdoc and Master’s student who conducted most of the data collection at the beamline. The second interview was about how the scientist managed the research data from the experiments, also bringing up questions on data transfer and data analysis. Interview and observation protocols are available upon request.

In total, five interviews were conducted with four individuals (the PIs were interviewed individually, whereas we interviewed the Master’s student and the postdoc in one session). The interviews lasted between 45 and 60 minutes, and were audio recorded to later be transcribed at a high level,

meaning that utterances were transcribed without taking notes of nuances provided by for example hesitancies, and paralinguistic aspects such as tone of voice. During the analysis, the transcripts were taken at face value, in accordance with the level of transcription we had chosen. A grounded theory approach was taken to the analysis, letting codes emerge from the data without an initial coding scheme guiding the analysis. In a larger-scale study generating larger amounts of data, a more rigorous data analysis approach would be taken. During the observations, we took notes on paper, and transferred them to digital form afterwards.

In addition to the data collection described above, we have studied different policies[3] and guidelines aimed at MAX IV's users, and had conversations with staff from the Controls & IT department, the beamline, and the User Office. This contextual information was gathered in order to better understand the conditions for the users of the beamline.

Findings

This section starts with a presentation of general findings, which is followed by sections on each of the three phases of the experiment process: before, during, and after the experiment.

Both PIs had visited NanoMAX once before and had experience from conducting experiments at other synchrotron facilities. The amount of experience of using synchrotron facilities among the members of the respective research groups varied, ranging from no experience to experience from several experiments at different facilities. The experimental data collected during beamtime were to be used by members of their research groups, including postdocs, PhD students and Master's students.

In all, both cases describe their experiences at NanoMAX as positive, especially the enthusiasm and helpfulness of the NanoMAX staff and the IT support staff - during the experiments as well as afterwards. Any glitches in the equipment and IT tools during their visits were attributed to the fact that the beamline is still under development, and were taken in stride. Both PIs expressed that they were very impressed with how much the MAX IV had evolved in a year since their last visits, in every aspect of the facility, from beamline instruments to tools, to the kitchen area. They also noted that there are many concurrent activities that the MAX IV staff need to focus on - simultaneously running the beamline *and* developing the beamline, including preparing for new users.

During the preparatory phase – before the visits

Both cases prepared their samples for the experiments most carefully, with the help of collaborators and from the NanoMAX staff. The sample preparations were also documented by the research groups with a typical laboratory routine for experiments. The study participants stressed the importance of arriving to the facility well prepared, with a sufficient number of samples in order to make use of the valuable beamtime. Both PIs had conducted experiments at NanoMAX a year previously, and therefore expected to have access to an electronic lab book to document the experiment process as they had the year before. They both expressed that the electronic lab book available at NanoMAX was appreciated during their last visits, and that such a tool was not available at other synchrotron facilities that they had visited. For the data analysis, both cases were preparing

[3] See https://www.maxiv.lu.se/users/user-policies/

software and scripts for continuing working with the data. These preparations were estimated to require quite some time of coding and scripting in different software and programming languages.

Digital User Office (DUO) - administration of the application and preparation process

The Digital User Office (DUO)[4] is an online tool used for administration of proposals and experiments. DUO covers the whole administrative process for users; from applying for beamtime to submitting experimental reports and giving visitor feedback. Furthermore, users are encouraged to register their publications based on their experiments through DUO. The study participants often referred to DUO during the interviews, so its role as a point of reference for planning and reporting experiments seems well understood. One of the study participants shared their experience of noticing during their visit that DUO was unavailable from inside MAX IV, and suggested that it become available in-house to enable checking details that are only available in DUO.

As stated, DUO focusses on administrative aspects of planning and reporting on experiments, so data and metadata are outside its scope. One piece of the research data management puzzle is "SciCat", a metadata catalogue under development, a collaboration between MAX IV, ESS[5] and PSI[6]. By linking the content in DUO and SciCat, users could potentially seamlessly access richer information about their experiment without having to use two different tools.

Suggestions

- Make DUO accessible to users during their visit (i.e. from inside the facility)
- Connect DUO and SciCat for a seamless user experience (long-term)

elogy – the electronic logbook

The electronic logbook elogy, developed in-house, assists the users with documenting their data capture during experiments. Both cases used elogy, and were satisfied with its functionality, stating that it was convenient to use.

After the experiments the users' notes in the logbook can be exported into a PDF file, enabling users to bring their notes home.

Three features were reportedly particularly appreciated:

- Linking between entries in the logbook, thus expanding the logging procedure from the sequential nature of a paper-based logbook to a dynamic experience.
- The logbook was accessible from several computers. At NanoMAX, it was available in two connecting rooms, allowing for distributed access to read the logbook. However, only one user at a time could edit the log. Some study participants described it as difficult to divide the editing responsibility among team members, i.e. deciding who should have the editing responsibility at specific times.
- Being able to export the logbook to PDF was very much appreciated, as it could easily be used in further work with the data. However, a clash of expectations was described by one of the study participants, as the static and sequential nature of the PDF document was in stark

[4] DUO, https://www.maxiv.lu.se/users/user-guide/
[5] ESS, https://europeanspallationsource.se/
[6] PSI, https://www.psi.ch/

contrast to the experienced dynamic nature of elogy.

The team in Case1 started off with an analogue lab book, taking their notes manually, but rather quickly moved to using only elogy. At first, they structured their notes chronologically, but after a while they switched to a structure by sample. It was the first time that the Master's student and postdoc used an electronic logbook, and they found it to be a very useful way of documenting their process. When elogy happened to freeze at one time, they promptly received help from the beamline staff.

The PI of Case2 stated that elogy had been improved and was more stable now, compared to last year's visit. Overall, they stated that elogy was very convenient to use, and particularly appreciated the possibility to add links into the logbook. A feature they would appreciated in the future would be to access elogy from home after the visit. They also suggested some basic user instructions for elogy.

Suggestions

- Make the logbook accessible to users after the visits, or downloadable in additional formats to make use of the experienced flexibility and dynamic nature of the logbook (e.g. linking). A version of the logbook could be locked and archived after the visit, allowing for another version to be accessed and used after the visit. This could contribute to the users' further work with data processing and analysis.
- Suggest some file structures that usually works for most users (e.g. sequential, by sample...)
- Post "Instructions for dummies" on a notice board in the control room. This could contribute to users becoming more self-sufficient over time.
- Make manuals available to users to study before the visit. This can be especially valuable for users who lack experience of synchrotron facilities.

ScanViewer – software to view and navigate image data

NanoMAX has developed a software called ScanViewer which allows users to view and navigate the image data in real-time during the experiments. Both cases used this software for navigation of samples and for recording phases of the experiment by copying and pasting images into the logbook, a feature they were very appreciative of.

ScanViewer was considered somewhat non-intuitive according to one of the study participants, who suggested that a basic user manual be made available, with pointers for new users, such as "don't use this command because the viewer will freeze" [Case1]. They also encountered some bugs that needs fixing, which would be conveyed to the beamline staff. Both research groups had tried to install the software at home for the data analysis, but had not managed to get it working.

During the processing and analysis phase – after the visits

Data transfer from MAX IV after the experiments

The data transfer was handled by the file transfer tool "Globus"[7]. Both cases had managed to transfer their raw data when we met them for the second interview, and reported that the transfer process had run smoothly. An improvement to Globus was suggested: a more intuitive user interface.

Data storage

Suggestion

- Develop a more intuitive user interface for Globus, e.g. a drag and drop interface

Both research groups store their data on their own servers. When asked about their thoughts on the time frame of the MAX IV data storage policy, both PIs were uncertain about the specific time limit. One of the study participants vaguely remembered reading somewhere about a 3-month time limit, but expected to receive an email from MAX IV well before data were to be deleted.

Data processing and analysis

Both cased stated that a lot of work remains after transferring data to their own servers, before they can start the data analysis. First, data processing has to be done, to make the data usable by various analysis tools and software, most developed in-house by each research group. The experiences of both cases were that the MAX IV staff was very engaged and helped out a lot in this process, both in selecting and recommending software for processing and analysis, as well as in converting data files into suitable formats for further processing and analysis. Case1 had a dialogue with the beamline staff before their visit regarding what data types their experiments would generate, and had been recommended using a software "PyMCA" for their analysis. In order to use PyMCA, the raw data files from the beamline had to be converted into a usable format, with which the beamline staff helped them. As Case2 experienced technical problems during their beamtime, they were not able to collect the amount of data planned. At the time of the second interview they had not yet started the data analysis, and had mostly checked that the data were readable. The plan for the analysis was to convert the data into a suitable format for analysis in the research group's own software.

Conclusion and suggestions for future studies

It is clear that MAX IV is under intense development, building beamline stations and accompanying instruments and equipment. The study participants expressed that they were well aware of this, and further expressed that they appreciated contributing to the development of MAX IV and NanoMAX. The results indicate that users could benefit from having more standard procedures and tools for processing and analysing data at hand, or at least suggestions of procedures.

Most of the experimental work was conducted by early career scientists, i.e. postdocs, PhD students, and Master's students. The principal investigators were heavily involved in the application process, in planning the experiments and in the scientific reporting after the visits. For junior scientists, the beamtime is a valuable learning experience about how to run experiments, and about the beamline station and the facility as a whole. Careful preparation and documentation of experiments is crucial in research. As digital tools for RDM, including documentation in electronic logbooks, are gaining traction, students and early career scientists could benefit from training and support in using

[7] Globus, https://www.globus.org/

available tools. This could be provided as part of the PhD education at their research institutions, and/or as support by MAX IV to users.

This case study is limited in time and scope, and we see several aspects that could be addressed more in-depth in future studies.

We followed the cases during a limited time span, i.e. a few months in total. This provides us with some insights into their RDM practices directly before and after the experiments, but misses out on the larger time scale. Follow-up interviews with users a few years after their visits could provide complementary information about their experiences, and how they ended up managing their data in the long-term (including whether they shared or published data). Another way of gathering more information about the users' needs for research data services would be to add a question about their experiences, allowing for suggestions for improvements, in the experimental report which is submitted by the users shortly after their beamtime.

More importantly, we suggest a large-scale study of users, broadening the scope to users with different experiences of synchrotron facilities, at different career stages, and from different research areas, to contribute more in-depth to the development of research data services. This could also contribute to the development of both general (supporting the commonalities across all, or most, user groups) and specialized (to accommodate specific needs) research data support and user support. A larger-scale study could also serve as a platform for knowledge exchange between beamlines and other organizational units within MAX IV working with research data in different stages of the research data life cycle.

An aspect that we discussed in the planning stage of the study, but which ended up peripheral, was how the institutional policies of the facility match the users' management of research data before, during, and after visiting MAX IV. The Experimental Data Policy and User Access Policy can provide points of reference between these perspectives, in enquiring how well the policies and the users' practices match; and if not, how any identified gaps may be overcome.