Data management policy AMOLF and ARCNL 2023-2028

Authored by the AMOLF/ARCNL Data Management team:
Bas Overvelde, Femius Koenderink, Oana Draghici, Jeroen van Zon, Peter Kraus, Roland Bliem, Marco Konijnenburg, Kristina Ganzinger, Cees van der Ven

Version: 2023-09-26

Executive summary

In 2018, AMOLF and ARCNL launched the implementation of a new Research Data Management policy (RDM), with the dual goal of improved accountability/ integrity, and ultimately FAIR Open Data. In the Dutch physics landscape, this policy has been ahead of policies at universities. Aided by the small size of AMOLF as compared to the disciplinary breadth of science faculties, we can provide specific and practical tools for the work floor.

The main pillars of our initially introduced policy are:

- Training for all scientists and technical support in data management tools, in the form of periodically held symposia with motivational and instructional talks.
- Data management plans for all researchers and projects. Junior researchers need to write, sign, and keep these plans up to date, to describe data handling and storage for all steps in the research process.
- Individually traceable paper logbooks with clear logbook instructions, centrally archived at AMOLF.
- "Data replication packages", defined as a structured and annotated collection of data files, logbook entries, processing scripts and blueprints to reconstruct the research result reported in the paper.

In Spring 2022, we have expanded this toolset to:

- Possibility of electronic logbooks as an alternative to paper logbooks. While groups are free to choose different electronic solutions, our policy provides clear requirements on its features and archiving.
- Training in improved coding practices, open data formats and use of version control.
- Offering an AMOLF portal at ZENODO to deposit data sets as Open Data.

We have the dual purpose of improving our scientific process accountability, from the viewpoint of research integrity and to move towards making all (valuable) data open access. Our philosophy for Open Data is that deposited data sets must be publication-quality datasets that are "as FAIR as possible". Our approach therefore has been to first internally further improve our research practice towards FAIR standards. In the first implementation phase, we focused on creating awareness and offering tools. Since 2021, we have as ambition that it is standard to deposit replication packages corresponding to each paper for which AMOLF has lead authorship in a closed institute archive, or as open data in ZENODO. In 2022 we have taken the next step, to stimulate researchers to make data sets open through ZENODO. Moreover, group leaders are asked to report on the implementation of this responsibility in our annual evaluation cycle. The number of deposited replication packages currently covers circa 25% of our papers. We expect this number to rise to circa 60%, bounded by the fact that for many collaborative papers, the data responsibility lies outside AMOLF.

Looking forward to 2024 and beyond, we will move to making OA deposition of data sets mandatory for all papers for which AMOLF or ARCNL is lead author, unless disclosure hampers international collaborations, IP creation, or working with industry. The most important tools for researchers to manage their data have been introduced in the previous period. For the coming period, particular attention will be paid to monitoring the implementation of our data management policy, training researchers when they start working at AMOLF or ARCNL, and stimulating researchers to ensure that their data is managed properly. Moreover, we will streamline the processes that occur “behind the scenes”, to ensure that we can effectively monitor the implementation of our data management policy and inform group leaders and researchers. With these implementations, we are reaching a steady state implementation of data management at AMOLF and ARCNL, and therefore we will also change the composition of our research group by placing more responsibility on junior researchers that span the AMOLF themes and expertise centers and ARCNL topics. Yet, we do see that the Dutch landscape for RDM is still changing, where new institutes are being founded and opportunities for exploring new projects related to RDM are appearing. We will ensure that suitable opportunities for AMOLF and ARCNL will be explored.

Table of content

1. Overarching policy ambitions

For the coming period 2023-2028 the focus will be to monitor the implementation of data management, to implement training for new researchers, and to make replication packages openly available. To achieve this, the DM team will:

- In 2024 start with the implementation to move towards making all (exceptions are possible) replication packages with AMOLF (or ARCNL) as lead author open access through ZENODO.
- Implement training during the onboarding of new research employees, similar to how, e.g., new staff is introduced to safety.
- Place emphasis on monitoring of implemented policy and tools, to determine improvements that can or need to be made for effective data management. For example, by implementing an *ad hoc* committee for reviewing paper and electronic logbooks, and by implementing an *ad hoc* committee that will review the quality of replication packages.
- Focus on implementing and improving new tools that are useful in monitoring the implementation of data management.
- Provide a more permanent solution for paper logbook storage.
- Introduction of a recurrent task list for the DM Team, that will be on the agenda of every meeting to check if specific tasks need to be performed. This recurrent task list will ensure that also new team members have insights into what needs to be done on a regular basis. Tasks that will be placed on this include, e.g., the updating of the DM plan, to request DM plans for grants, and to list open access replication packages during the staff meeting)
- Change the composition of the data management team, to have a stronger representation of researchers from AMOLF and ARCNL that use the data management tools daily, and support staff that works on the implementation and monitoring of data management. The composition of the data management team will also better reflect the different themes and/or expertise centers.
- Maintain the same yearly budget of 20 k€/year for the coming period, to allow flexibility to explore new tools to, e.g., train new researchers.

How to read this document?

With the initial data management policy set out in 2017, we have set the basis for data management at AMOLF and ARCNL. This document should be seen considering the previously introduced policy and provides an overview of (practical) considerations that need to be made in the coming period (2023-2028). It does not provide full details on the exact implementation details, but more on the general lines and next steps that need to be taken.

2. Implemented data management tools

Since the introduction of the previous Data Management Policy covering the period 2017-2021, at AMOLF and ARCNL several data management tools have been implemented to improve and streamline data management. The focus was mainly on improving internal data management processes and awareness. The first tools for researchers have been adopted on November 1 2019, and covered project-level data management plans (section 2A), paper log books (section 2B) and replications packages (section 2C). Additionally, on March 31 2022 we have also implemented best practices for Data Management plans for grants (section 2A), making data open access through ZENODO (section 2D) and making use of version control tools (section 2E). The best practices for data management and the implemented tools at AMOLF and ARCNL have been shared during two training events on 23/24 November 2019 and 31 March 2022, where it was mandatory for all researchers to join. All information regarding data management at AMOLF and ARCNL can be found on the intranet:

https://intranet.amolf.nl/dms/

2A. Data management plans

Internal data management plans

Each AMOLF/ARCNL project is covered by a Data Management plan (DM plan) describing the data acquisition, processing workflow and storage during and after the project. DM plans cover all data of the project, including logbooks, stored data on computers during and after the project, and the data that is provided in replication packages belonging to published articles. DM plans are therefore a tool to make an inventory of data storage needs and ambitions before a project starts, and to make explicit the way in which a research group stores and shares data.

The granularity of DM plans is set by group leaders to for instance a per project or per collaboration. The DM plans are written and updated by the junior researchers executing the project to which the plan applies. A DM plan should be created at the beginning of a project, ideally at the start of the new junior researcher's employment. DM plans are not carved in stone, they can grow and adapt during the project.

Existing template, etc.

To make writing DM plans a structured and manageable process, AMOLF/ARCNL provides two documents: (1) a template, and (2) an example of a DM plan. Note that much of the DM plan has already been provided in the template, as much of the information is institute wide. The researcher should carefully read these policies to memorize themselves with the AMOLF/ARCNL policies, the DM plan thus provides instructions for the researchers on how to manage data during their time at AMOLF/ARCNL. The rest of the template document provides a standardized questionnaire that needs to be filled in with group-wide and project-specific information. Typically, there is a large overlap between DM plans within a research group due to the use of similar equipment, fabrication approaches, software, etc. Therefore, it is advised to also consult with other group members and your group leader.

State of implementation

To determine if researchers are familiar with DM plans, in the most recent training event we asked the participants (119 total) if they have written a DM plan and if they keep the DM plan up to date. We found that 65,5% never has written a DM plan, 21,8% only wrote a DM plan once, 8,4% wrote a DM plan for each project, and only 4.2% of the participants was actively updating the DM plans.

From an implementation perspective, policies for writing the DM plan, including the provided templates, are fully implemented. Still, as the DM plan should contain the most up-to-date policy with the updated policy document, introduction of new tools, etc., the DM plan needs to be kept up to date.

To be improved/implemented

To ensure researchers know how and when to write DM plans, we need to ensure that they will get properly trained the moment they start at AMOLF or ARCNL. For that, it is important that training will be provided and, similar to safety instructions, evaluated if they followed the training. Additionally, clear responsibilities should be defined to introduce additional moments where the implementation of DM plans are checked. The general AMOLF/ARCNL philosophy is that the group leaders bear responsibility for implementation of policies in their group on basis of mandate by the director. This means that group leaders are free to determine how their group formulates its DM plans, and are responsible for overseeing that there are up-to-date and consistent DM plans, which junior scientists update and adhere to. The annual evaluation cycle of each junior scientist invites the group leader and junior researcher to jointly review and update the DM plans, as a periodic failsafe for what ideally is a continuous awareness of DM practices. In the annual P&C cycle of the director with the group leaders, group leaders should self-report on how they implemented DM-plans in their group and their experiences. An overview of these responsibilities can be found in section 4C.

Besides the aforementioned responsibilities, the role of the DM team and of ICT is to provide templates, instruction, and a dedicated storage location for the DM plans. With the introduction of new tools and changing policies we need to ensure that the DM plan provides the most up-to-date information for researchers. We will therefore add a task to the recurrent task list for the DM team, where we will yearly check if the DM plan template needs to be updated and do so accordingly.

Data management plans for grants

Existing strategy

Funding agents such as NWO and the EU are requesting DM plans to be submitted after they have awarded your grant, and before the grant starts. We have collected examples from previous awarded grants and grant applications to aid this process, so that group leaders can base their own data management plans on examples from colleagues. You can find these examples on the SUN repository:

sun.amolf.nl\support\Datamanagement\DM_Plans

Moreover, as of 2020 NWO asks that you assert that your template has been discussed with a data steward/responsible person. Since we do not have a dedicated data steward, this mandatory consultation is with a member of the data management team. Group leaders can send an email to datamanagement@amolf.nl or datamanagement@arcnl.nl to request such consultation, the chair of the data management team is acting as consultant. We will continue to update the repository of data management plans used for grants. Note that obviously projects with IP constraints merit extra attention.

To be improved/implemented

For the DM Plans for group leaders, we need to ensure that the list of DM plans belonging to successful proposals (and the DM plan template for projects) needs to be kept up to date, such that the examples used for writing new DM plans are still relevant. We will therefore add a task to the recurrent task list, which includes a yearly email sent to all group leaders to request DM plans for successful proposals, such that these can be added to the repository on the intranet.

2B. Logbooks

The use of a logbook to record ideas, inventions, experimentation records, and observations is an important part of any scientific research. A successful logbook enables proof of the quality and integrity of research data and enables others to reproduce the data to achieve the same result. Another important purpose of a logbook is to support the documentation of work that may be patentable. It is therefore necessary to provide clear, concise, and chronological entries with specific data. We provide researchers with guidelines to help them create more efficient and accurate logbook entries, where guidelines are available for both paper and electronic logbooks.

Paper logbooks

Existing logbook and workflow

The Librarian ensures the acquisition, registration, monitoring and storage of standard paper logbooks. At AMOLF, researchers will get their logbook from the librarian. At ARCNL, the hand-out of paper logbooks is done by the ARCNL secretariat. If a logbook is completed, further paper logbooks can be obtained at the same location. Each paper logbook is registered with the librarian at issuance in a database containing the institute, the issuing date, the name of the person using it and, after completion, the end date. All registered paper logbooks can be consulted at:

http://alida4.amolf.nl/find/logbooks

On the last working day of the researcher at AMOLF/ARCNL they must return all their logbooks to the librarian where they will be archived. Exemptions can be made in consultation with the group leader. Logbooks and the intellectual property contained therein belong to AMOLF/ARCNL and should remain permanently at AMOLF/ARCNL. Each logbook must be registered. Pages of the logbook may be copied after consultation with the group leader.

State of implementation

Currently practically all researchers using paper to log their research indeed use the AMOLF/ARCNL - supplied logbooks. Quality-control during use is the responsibility of the group leaders. Logbooks are to be handed in for archiving when students leave, and often retained somewhat beyond that by groups, e.g., for supporting the writing of papers. It is therefore difficult to assess the "coverage" of the archive. For instance, for AMOLF there are currently (March 2023) 170 logbooks in use, over 130 in archive, and circa 80 that have been used but were not archived. Of those not handed in, ca. 75% belong to internship students (projects not resulting in papers), and 25% to recently departed students. ARCNL has 80 logbooks in active use, circa 30 in archive, and circa 20 not returned to archive. The relatively smaller number of logbooks that have been archived is likely because we have only recently reached "steady state". The yearly P&C cycle for each group queries the group leader on all the unarchived logbooks.

To be improved/implemented

While the necessary information for researchers to properly use logbooks is available on the intranet and will also be provided in a future training module that we will develop (section 4A), it is important that group leaders take their quality assurance role serious. For overall quality improvement we will have an *ad hoc* committee review paper (and electronic) logbooks in 2024 to collect recommendations for improvement.

Furthermore, we should ensure that the archiving capacity is adequate for the duration that paper logbooks should be stored. This duration is at the moment still not defined and might depend on the practical implementation of the archive. At the moment, we estimate that logbooks place the following space requirements:

What	**Shelf space in meters**	**Where**	**For who**
Boxes of unused stock	3 m	Hallway in front of librarian office	AMOLF
	2.3 m	Hallway in front of librarian office	ARCNL
Running stock ready for hand out	3 m	AMOLF librarian office	AMOLF
	2.3 m	ARCNL secretariat	ARCNL
Archiving	2.5 m / year	Hallway in front of librarian office	AMOLF
	1.2 m / year	Hallway in front of librarian office	ARCNL

We note that a 10 year archive would imply 25 m / 12 m of shelf space for AMOLF/ARCNL, which is equivalent to 6 respectively 3 bookcases. In the coming period a more permanent solution for archiving should be defined. Questions that need to be answered are: Where and how will we store logbooks?; Do we want to digitalize logbooks after use?; How long do paper logbooks need to be stored?

Electronic logbooks

Existing policy

The use of an electronic logbook is accepted at the discretion of the group leader if it complies with the provided guidelines on the intranet. If researchers are using an electronic notebook, they have to register it with the following tool by using your login AMOLF/ARCNL credentials:

http://alida4.amolf.nl/user/logbooks

It is important that the integrity of record is maintained (and not altered during the logbook keeping). We therefore recommend archiving the online logbook (e.g. in pdf format) at a tamper-proof location on the server. For now the solution is the following: after contacting the datamanagement team, your group can get an ELN-archive folder in the group folder on the server, where the whole group can save documents, which cannot be edited afterwards (i.e. the whole group has saving, but no editing rights). Logbook files should be archived with a frequency of two to four weeks to ensure the integrity of record. This routine also ensures the logbook cannot be made illegible.

On the last working day of the researcher at AMOLF/ARCNL the researchers need to ensure that an electronic copy of the full logbook is available on the AMOLF server, at the same location where copies of the logbook were stored.

State of implementation

Since the introduction of this policy, we see that researchers have started to use ELNs. Examples of registered electronic logbooks that are being used are: Microsoft OneNote, Benchling, reMarkable, elabFTW or group Wiki pages. We see that more people are starting to use ELNs. The long-term storage is kept on the ‘sun’ server, which is supported by the AMOLF/ARCNL ICT team, folders have been generated on group drives. An archive folder can be requested via ICT that has access (reading and writing) rights that ensure the integrity of records. Besides some technical hiccups this systems in place seems ok, although it is difficult to monitor how well logbooks are backed up at the moment, because the storage is only accessible by the groups. Instructions regarding the procedure of getting an archive folder are given in the ELN instructions sheet (intranet). While the stored logbooks allow tracing who uses which logbook, we do not have a centralized overview via a registration system yet of how many researchers are using ELNs, and which type of ELN they are using.

To be improved/implemented

As mentioned researchers can choose their own ELN platform as long as a commonly accessible format (pdf) is archived on the server. The DM team will keep monitoring if this solution is viable, or if an institute wide ELN solution has to be offered. Moreover, the registration system for logbooks is in place, but needs to be advertised more, which would in turn give a centralized overview which electronic logbooks are in use. This is connected to training in DM, which will be a focus in the coming period.

The same holds true for storing the ELN on the SUN drive, and the importance of keeping integrity of record should be stressed. The advice is to automate the monitoring of registering e-logbooks in ALIDA, and the monitoring of archiving, if this can be practically done. The plurality of e-logbooks solutions that we allow might make this difficult.

For overall quality improvement we will have an *ad hoc* committee review electronic logbooks (similar to the review of paper logbooks) in 2024 to collect recommendations for improvement.

2C. Replication packages

The goal of a replication package (RP) is to provide a minimal yet as complete as possible set of information by which an interested third party could in principle independently replicate the results of a paper published in a peer reviewed journal. This material will typically include data, protocols, and analysis scripts organized in a logical manner and provided with adequate metadata.

Existing policy

A RP does not need to play the role of the final storage location of the complete research data, but rather provides the minimal sufficient set of data and methods for replication. Elements should be added to it on a "need-to-know" basis in the context of the specific paper and include raw data in so far as this is reasonable from a storage and handling perspective. AMOLF/ARCNL will create and store an RP if the most significant part of the work was performed at AMOLF/ARCNL, as evidenced by the affiliation of the first author, the corresponding author and/or the last author.

Each RP will be stored in a separate directory. The structure of an RP is paper centric. In this structure, the figures and tables of the published paper are the starting point, and all the collected material is tied to these items. Note that this structure still allows for a lot of freedom in organizing the material. To facilitate understanding of the structure and contents of the RP metadata README files should also be supplied.

One of the authors designated as submitting author by the AMOLF/ARCNL group leaders involved oversee and submits the RP when the final version of the paper is officially published, i.e., either online or in print, but not on e.g., ArXiv. Touchstone is that a DOI has been assigned. The submission involves two steps. First, the RP is prepared in a directory. The current limit is 2GB. The finalized RP is copied to a folder on the central storage environment of the group to which the designated author belongs. Second, the submitting author fills in an online form that can be found on the intranet.

After these steps have been successfully completed the ICT department will review your replication package and if everything is ok, they will move it to the replication package repository.

State of implementation

In the last years we have seen an increasing number of people preparing and submitting replication packages for articles, indicating that the proposed method can work. Yet, we still see that for most articles there is no corresponding replication package, which can be expected when introducing new policies, as there is a delay between publication of articles and the time of research (when data should be stored). For example, for articles published in 2022 for AMOLF and ARCNL roughly 30% a replication package exists (23/73 for AMOLF, 13/41 for ARCNL). Still, given that the policy was introduced approximately 4 years ago, we expect that the current level of replication package submissions has reached "steady state".

To be improved/implemented

One potential way to increase the number of replication packages that are submitted for published paper is to send researchers (and group leaders) reminders. In current implementation of the RP policy there is no (automated) system to screen if publications have associated RPs, and no (manual or automated) system to then produce reminder e-mails. This requires tooling from ICT.

Furthermore, at the moment there is no automated RP-submit procedure: the web-form leads to manual intervention from ICT to move the RP to archive. This requires tooling for automation, to improve monitoring but also allow more regular updates to group leaders and researchers in relation to the required submission of RPs. To improve diligence in submitting RPs, the monthly staff reporting of "Publications" will be replaced by a report that lists publications of the last 3 months in a tabulated format *including* an indication of whether RP packages are present, and if they are published open access on Zenodo (see section 2D).

Finally, to determine the success of the RP policy and the quality of the RPs submitted, we propose the following:

- Quality: In 2023, an *ad hoc* committee is composed to review the deposited RP packages and to return recommendations for quality improvement. At AMOLF, this could for example be done within the expertise centers.
- Quantity: the number of RP packages per year can be reviewed at the annual output document stage – which is shortly after the annual P & C cycle that serves as final reminder to submit RPs (with a suggestion to send the list of replication packages one month prior to the P & C meeting to inform the group leader). The number of RP packages can be judged against an expected "coverage" (to be determined), considering the historical percentage of papers where AMOLF/ARCNL is lead author.

2D. ZENODO

ZENODO is a free data-repository hosted by CERN. The primary purpose of ZENODO is as an archive to deposit high-quality data sets in, which are then:

(1) Findable through a unique persistent identifier or Digital Object Identifier (DOI) and associated searchable metadata.
(2) Secured for a long term (20 years) in a form that allows no alteration.
(3) Accessible through making data “Open Access”.

Thus, ZENODO is mainly a data repository for Open Data. However, ZENODO also provides embargoed, restricted, or closed access deposition. In this case the DOI and metadata are still findable, but the data itself is not openly accessible.

Existing policy

AMOLF/ARCNL designate ZENODO as repository of choice to make datasets associated to publications, and other publication-quality data sets “Open Access”. Up till now AMOLF/ARCNL did not require all data to be open access, only to provide a platform in case researchers want to make their published work open access. One of the following two scenarios most likely apply:

(1) A paper has been accepted for publication in a scientific journal, and so you have prepared a replication package associated with the publication. The authors wish to make it Open Access. The best moment to post the data set is at the proof stage of your paper. The ZENODO repository can then link explicitly to the paper DOI, and you can still insert the data DOI in the paper proofs, so that the published paper references the data.
(2) Already at review or preprint stage, reviewers and readers need to have access to the data associated to the paper. A solution is to post the data on ZENODO before submitting to the journal. The submitted preprint can point to the data DOI. ZENODO allows versioning, which accommodates revisions to the data in review rounds. A variation of this scenario is to post the preprint on arXiv / ChemRxiv / bioRxiv and the data on ZENODO, with cross referencing, and enter the review process with preprint and data already public (not all, but most, journals allow preprint server posting).

Note that ZENODO does not replace the requirement to also store a replication package internally on the server. If you wish to store a replication package but keep it closed, there is no need to put it on ZENODO and only the internal replication package procedure applies.

Data is deposited under the responsibility of the senior author (group leader). Once the data is published, it is immutable. However, ZENODO does allow metadata modification and also versioning. Technically the data is deposited from a ZENODO account holder and use of versioning / changes to the metadata is accessible only to the original depositor account. It is hence preferred that the group leader submits, at least when versioning is anticipated to be relevant (mainly scenario B above).

While the ZENODO upload process is easy, we recommend everyone to consult the manual on intranet. This is to maintain an organized AMOLF and ARCNL curated community for facilitating the linking of data DOIs to the institute repository of publications, and to ensure the use of ORCID identifiers to link datasets to the track record of all authors.

State of implementation

Currently, the number of groups at AMOLF and ARCNL publishing research-related content on Zenodo is small. A search on Zenodo (performed on 03-03-2023) yields a total of 32 items (18 data sets, 8 code files, 5 journal articles, and 1 design file) for AMOLF and 4 items (1 data set, 2 journal articles, 1 report) for ARCNL. The type of uploaded content, the list of included authors, and the format of the title and description differ significantly between the items. Only a few of these results are linked to the AMOLF Zenodo Community; none for ARCNL. Nevertheless, the items are findable using the search terms AMOLF, ARCNL, and "Advanced Research Center for Nanolithography".

Ambition

Ultimately AMOLF/ARCNL will require all data sets to be "Open Access", unless a clear reason not to (e.g., IP reasons) is identified. Up till now (2022), use of ZENODO was optional, but institute-specific policies to increase the number of openly accessible datasets are expected to take shape in early 2024. This will of course go hand in hand with an increase in the percentage of replication packages that will be prepared for each published article (see Section 2C)

To be improved/implemented

The data management team will support the increased use of Zenodo by providing additional information and tools facilitating open access data storage. Small extensions to the guidelines on replication packages (RPs) and on the use of Zenodo will clarify the connections and differences between RPs and open access data. The RPs submitted to the institute repository will not automatically be published on Zenodo. They can be interpreted as a structured collection of the minimum amount of data that is reasonable to share via open access storage solutions. For RPs and open access data, the responsibility for the quality, accessibility, interoperability, and reusability (see FAIR principles) of the datasets lies with the group leaders supervising the research and the author uploading the data. The findability of the datasets on Zenodo will be ensured by further improving the guidelines for the submission process and emphasizing the relevance of linking to the institute environments on Zenodo, and by providing guidelines for the required metadata that is published on Zenodo together with the RPs. Specifically, guidelines for the preferred format of title, author list, and description of the Zenodo items will be provided and an explanation of the Zenodo environment will be included.

An increasing number of funding agencies and publishers requires statements on open data for applications, articles, and related documents. The data management team plans to provide templates, examples, and standardized statements on Open Data that underline the AMOLF/ARCNL policy and that can be directly used by AMOLF/ARCNL staff in e.g. funding applications and article submissions. Of course, it should be noted that in some cases it is not possible to publish data open access, and to identify these cases the DM team will provide a decision-making tree.

To monitor the increase of openly available replication packages on Zenodo we will:

- Ensure that ZENODO submissions are associated to the AMOLF/ARCNL community for the library to be automatically informed. (Currently not all ZENODO submitted packages are linked to AMOLF or ARCNL).
- Yearly map the number of available replication packages by placing this as a task on the recurrent task list of the Data Management team.
- Indicate on the publication list that is submitted to the staff meetings which publication contains a replication package, and if that replication package has been published on Zenodo.
- During the P&C cycle indicate which articles don't have an openly available replication package.

2E. Version control tools

Version control tools are software that allows systematic tracking of changes to files, and typically used for tracking changes while writing code, articles, protocols, etc. AMOLF/ARCNL provide support for Git and Subversion, two main tools used for version control.

Existing policy

At the moment, there is no direct policy within AMOLF/ARCNL to use version control. However, it is advised to do so to maintain a proper record of written code. To facilitate the use of version control tools, we provide:

(1) A brief presentation introducing the key ideas behind version control, available on the intranet.
(2) Basic tutorials to get started with GIT or Subversion within the context of AMOLF/ARCNL, available on the intranet.
(3) Both a Git and Subversion server for AMOLF/ARCNL, to facilitate sharing and collaborating on code within each institute.

State of implementation

On the Git server (git.amolf.nl), there are about 90 registered users, with about 80 code repositories. On the Subversion server (subversion.amolf.nl), there are about 30 registered users. In general, users are divided over both the support and scientific groups. The number of people using version control is likely higher, as an unknown number of people instead use public repositories, like GitHub. Both the Git and Subversion server are intended purely for internal use and don't facilitate sharing code outside of AMOLF/ARCNL. However, once published, repositories can be easily transferred to public repositories, like Github, where they are accessible for outside users.

Ambition

We intend to make the barrier for users to use version control as low as possible, by providing tutorials and easy access to internal and external repositories. At the same time, we recognize that there is a range of people writing code, for an equally broad range of applications. Using version control software can be challenging for users that have limited experience with coding. Moreover, for simple data analysis scripts that are written and adapted during a research project it is sufficient to only store the final version used to generate published data. Instead, version control software is very important for complex software, that is maintained and used by multiple users over a prolonged period of time. Examples are simulation software, acquisition software or data analysis packages. Typically, version control software is already used by research and support groups that work on such complex software projects, although this can likely still be improved among research groups with less experience working on such projects.

To stimulate the use of version control tools, the internally organized Python Workshop' also promotes the use of a versioning system. For most people this is a change of habit, and is slow to change.

2G. Requirements on offered storage

Existing policy

At the moment, research data during the duration of a project is stored on the SUN file system. This data contains 'raw' research data, such as unprocessed microscopy data or measured signals, but also scripts used to analyse the data.

The SUN is an SUN Oracle storage appliance. The system replicates data from Main storage (SUN) to replication storage (MOON). Snapshots are taken to provide history and recovery of data.

State of implementation

The amount of data is growing linearly. The total capacity is monitored by the ICT team. The individual groups have a quota which prevents the system from running out of space. Due to the quota, some groups face issues with storing larger datasets. Datasets need to be kept on the storage system for 10 years after publication. Consequently, in the current implementation, large, published data sets still occupy SUN storage for a long time, even if the data itself is no longer used, exacerbating existing problems of limited storage space.

Ambition

We intend to create a 2 tier storage system.

- Hot data on main storage (tier1).
- Cold data to be archived to less costly storage (tier2).

The 2-tier storage system is planned to be available in 2023. The 2nd tier can be outside AMOLF/ARCNL on cloud-based solutions. It is important to note that the main implementation considerations for storage fall (and will remain) outside the scope of the Data Management team, and are the responsibility of ICT.

3. Data management landscape

3A. NWO's Open Science and research data management policy

With the international trend to promote open science, NWO has introduced its research data management policy in 2016 [https://www.nwo.nl/en/research-data-management] (in addition to striving for all publications it funds to become openly available). This policy aims to make research data that is generated as part of an NWO funded project openly available, and as FAIR (findable, accessible, interoperable, and reusable) as possible, [https://www.nwo.nl/en/open-science] given that research data management is part of good research practice. The focus of NWO is both on internal (e.g. safe storage and careful curation) and external (available for reuse as widely and early as possible) processes.

NWO puts the following expectations on researchers: (1) carefully manage all research data, (2) preserve data for at least ten years, (3) share research data that underlie research publications alongside those publications, (4) deposit research data in a trusted repository following as FAIR as possible. To achieve this, researchers (for AMOLF/ARCNL typically the group leaders) are asked to answer a few questions in the Data Management section of proposals, and if successful, should submit a Data Management Plan to NWO.

3B. Changing Dutch landscape

With the changing Dutch landscape it is difficult to make a precise overview of the most important players for Data Management in the Netherlands. This section given an overview of the most important players for AMOLF/ARCNL medio 2023.

The NWO policy has received structural investment and kickstarter funding in 2019 through the "Implementation Plan Investment Digital Research Infrastructure", with a specific focus on digitalization. [https://www.nwo.nl/en/researchprogrammes/implementation-plan-investments-digital-research-infrastructure] Within this program there are four funding lines, namely local DCCs, thematic DCCs, investments in eScience, and Computing facilities. The local and thematic DCCs are particularly interesting from a Data Management perspective. Besides these investments, in 2023 we have also seen the introduction of the Open Science NL group within NWO.

Local DCC for NWO-I

Unlike most universities, AMOLF and ARCNL do not have an onsite local Digital Competence Center (DCC) with dedicated staff that support researchers in e.g. making their research data FAIR or improving research software and computing practices. Therefore, AMOLF and ARCNL have been part of an initiative to set up a local DCC that supports all NWO-I institutes, through initial investments within the Implementation Plan Investments Digital Research Infrastructure. This local DCC (1fte) has been in place since 2021, and focusses on generating knowledge (training), advice, awareness and representation of the institutes at both a national and international level. The local DCC of NWO-I focused mainly on joining the Carpentries, [https://carpentries.org/] within which researchers can prepare training in digital competences for other researchers (e.g. a training has been developed for the use of python and version control tools).

Thematic DCCs

In the spring of 2022 the thematic DCC (TDCC) network organization has started. The purpose of TDCCs is to provide means and funding for researchers to collaborate across institutions on topics related to open science. Three TDCCs have been defined, where the TDCC for the domain of Natural and Engineering Sciences (NES) is most relevant to AMOLF and ARCNL. Within the next 10 years, NWO will invest 2.4 Million euros annually to support these TDCCs and funding schemes for researchers, to support concrete projects within digital needs that have been identified.

Open Science NL (regieorgaan open science)

In the spring of 2023, NWO officially launched Open Science NL. Open Science NL is based on the initially launched National Programme on Opens Science (NPOS), and will replace this older initiative that launched in 2017. The aim of Open Science NL is to stimulate and speed up the transition of open science in the Netherlands. The organisations (which is housed within NWO) will identify, prioritize and finance projects related to open science. It will monitor the progress of open science in the Netherlands, and with identify critical points in the open science agenda. Moreover, it will provide a forum to share knowledge and to stimulate best practices. It is likely that Open Science NL will become the main player that sets the Dutch open science agenda in the coming years.

LCRDM

The National Coordination Points Research Data Management (LCRDM) is a national network of experts in the field of research data management. Where Open Science NL is mainly focusing on policy and the open science agenda, LCRDM aims to be the link between policy and solution. The LCRDM is housed within SURF. An important task that the LCRDM has taken up is for example the organisation of the contact point for Local DCC‘s, and training of people that are new in the job of research data. Moreover, the LCRDM has an advisory group consisting of people from SURF, universities and applied universities that aims at directing the national cooperation in research data management.

SURF, DANS and 4TU.ResearchData

Besides the aforementioned institutes that focus more on policy and bridging between policy and implementation, there are also several institutes that focus on providing support and tools for data management, such as data storage and open access publication of research data. Examples of these institutes are SURF, DANS and 4TU.Researchdata.

3C. Funding agency DM plans, rules

Besides the national policy that is reaching AMOLF and ARCNL through various instances, from a more practical perspective researchers will directly come into contact with research data management when applying for funding and when submitting articles to journals. Also here we see that the landscape and expectations are constantly changing.

NWO is now expecting a data management plan for every project that gets funded (and already during the submission of the proposal some questions have to be answered). For projects funded by the European Union, there is not strictly this requirement as you can opt out for writing a data management plan. Yet, this is currently a pilot, and we also expect that here more strict requirements will follow for funded projects. How each data management plan will look will likely depend on the funding agency. This is why we also chose to share examples of data management plans for grants (see section 2A) to allow for flexibility, rather than having a specific format for a data management plan provided by AMOLF/ARCNL.

We also observe a trend for submitted papers that data is becoming a more integral part of the process, either after acceptance (e.g. by submitting data to a repository), but also already during submission of the articles where authors are expected to submit all their data and code already for review. This makes it essential to have proper data management tools in place, to ensure that researchers at AMOLF and ARCNL are ready for an increasing requirement on data management and software development.

4. Training, monitoring and responsibilities

4A. Training so far

In the past we have organized *ad hoc* mandatory events for the entirety of AMOLF and ARCNL to introduce researchers to our data management policy. So far we have had two events, the first on 23 and 24 November 2019 to introduce the new data management policy and tools related to paper logbooks, data management plans and replication packages, and the second event on 31 March 2022 to introduce new researchers to the AMOLF/ARCNL data management policy, but also to introduce new tools to all researchers, which include the use of electronic logbooks, use of Zenodo and version control tools. These events actually served the triple purpose of (1) training, (2) motivation and inspiration, and (3) collecting feedback. For new employees, the intranet website contains instructions as video and documents from these events, which we expect group leaders to point employees to during onboarding.

Our steady state vision

While we had two data management events, these events are not sufficient to directly train new researchers when they enter AMOLF/ARCNL. As such, training is done mostly through the available information provided on the intranet page. Training is therefore dependent on the interest and commitment of the researchers. In general, we want to improve the awareness of the importance of data management in doing research, which should therefore be also highlighted during onboarding.

In the future, we propose to systematically tackle the training aspect for new researchers (interns, PhDs, PostDocs and Group leaders) through an e-learning system. The training materials should also be available as a resource for experienced users. The e-learning system could be the existing LabServant system or a new chosen one (Moodle), that suits not only this training but also others like general safety, laser safety, knowledge safety, inclusivity, etc. While ultimately group leaders are responsible that their group members are trained in data management, a workflow similar to the current ARBO training can automate the monitoring. The e-learning system should have an (automated) monitoring tool to track attendance and to follow up if taking the questionnaire is taking too long. We expect this new training approach to take shape in early 2024.

Besides training researchers during onboarding, the data management team will continue to introduce new tools if deemed necessary for improved data management. To make people aware of these new tools, and to keep data management part of the agenda and for feedback, we will keep organizing data management events in the future. By recording these events, we will also be able to generate the required information for training that is provided on the intranet.

4B. Staff engagement activities

Besides the training during the onboarding of new employees and the events organized when introducing new tools, we will also implement additional measures to keep data management part of the agenda. To do so, we have already introduced data management to the agenda of the annual Planning & Control meetings for group leaders, updated the annual performance evaluations for PhD students and PostDocs, and updated the to do list when exiting AMOLF/ARCNL. In addition, for coming period we plan to present best and worst practices during staff meetings, pay additional attention to Data Management during the Planning & Control meetings, and provide more information on the availability of Open Access replication packages in staff meetings. With the addition of junior researchers in the data management team (See section 4D) we will also explore other engagement activities that will motivate junior researchers to continue to perform proper data management. Examples include a yearly price for the best replication package, data set, etc.,

4C. Responsibilities

With the introduction of data management policies there should also be a clear overview of the responsibilities at all levels of the organization to ensure proper data management. This section gives an overview of the responsibilities related to using and monitoring the proper use of logbooks (both paper and electronic), data management plans, and replication packages (including open access through ZENODO). Moreover, we also indicate the responsible to ensure that new researchers are trained in data management.

Paper logbooks

Paper logbooks follow the following chart of monitoring responsibilities

Who	**What**	**When**
AMOLF: Librarian ARCNL: secretariat	Hand-out Register in ALIDA database	At contract start, and as need requires
Group leader	Use, and quality	Daily supervision Annual FUBO review
Librarian	Archiving Registering as archived	At contract end, part of “Leaving institute” checklist
Director	Monitor archiving	P&C cycle, based on lists provided by library

Electronic logbooks

Electronic logbooks follow the following chart of monitoring responsibilities:

Who	What	When
Researcher [PhD, PD, intern, technician, or GL]	Self-registers in ALIDA database	After GL has approved the proposed e-logbook as compliant with the e-logbook requirements.
Librarian	Monitors through ALIDA “list-of-employment" to spot if researchers appear to not have registered. Follows up with GL	Quarterly
Group leader	Quality Archiving	Throughout. The (at least) weekly archiving on sun as (watermarked) PDFs is imperative.
Director	Monitoring of archiving, use.	P & C cycle., GL reports.

Data management plans

Responsibilities for the AMOLF-internal data management plans are hierarchically distributed as follows:

Who	What	When
Junior scientist	Writes, adapts, signs, follows DM plan	Integral part of research
Group leader	Verifies with junior scientist that DM plan is up to date and followed. The group leader is responsible for/oversees the presence, maintenance and adherence to DM plans.	Annual FUBO cycle (form item)
Institute director	Verify with group leader that group has up-to-date DM plans	Annual P&C cycle (agenda item)
DM team	Assists with template, instruction	
ICT	Sun has DM plan directories	

Replication packages

This subsection considers the responsibilities and monitoring of replication packages that AMOLF/ARCNL requires for storage in its internal archive, or alternatively on ZENODO, for each paper in which AMOLF/ARCNL is the lead authoring institute. The responsibilities and monitoring cycle are as follows:

Who	What	When
Leading author, under responsibility of group leader	Composes and internally submits RP package If existing, reports ZENODO DOI	At paper publication, with a 1 month grace period
Library / head DM team	Check presence RP / ZENODO, and for ZENODO the AMOLF/ARCNL community association and send (automated) reminders to group leader	1 month and 3 months after paper publication
Institute director	Verify with group leader that group is up to date with RPs for the past year	Annual P&C cycle Agenda item, with library list for reference
DM team	Instruction on request Periodic review of process flow Periodic review of RP quality	
ICT	Functional (automated) RP-submit procedure Tooling for library to report on RP presence, reminders	

This scheme follows the general AMOLF/ARCNL philosophy that the group leaders bear responsibility for group output, with the possibility to mandate the execution to the junior scientists.

ZENODO

Once ZENODO is a standing policy, the same hierarchical flow diagram of responsibilities for replication packages can be used. A crucial point is that such a policy will be of the form “data is open [ZENODO]...unless”, where group leader can make a motivated decision to place the data in the local repository as replication package should there be valid reasons for an embargo. At the time of publication this leaves the flow chart of actions intact. At the P&C cycle, the rationale for embargoes can be reviewed by the director.

Training

The proposed e-learning system that will be implemented in early 2024 should have a monitoring tool to track attendance and to follow up if taking the questionnaire is taking too long. Proposed moments and actors are defined in the following table:

Who	What	When
E-learning admin	Enroll new employees.	On arrival new person.
E-learning admin	If not completed, send reminder to employee.	One month after enrollment.
E-learning admin	If not completed, send notification to employee and groupleader.	Two months after enrollment.
E-learning admin	If not completed, send notification to employee, groupleader and director.	Three months after enrollment.

For ARBO trainings, this workflow is currently done by hand by the ARBO coordinator. The current system does not allow automated messages or produce the info needed to automate this and reduce manual actions. Since a training system is now being considered for multiple AMOLF/ARCNL policies, we could envision a more central role, for instance HR, or administration/reception.

4D. Resources

Finances

For previous period a budget of 20k€/year was provided to implement data management at AMOLF and ARCNL. This budget has been sufficient for previous period, and most of the budget that has been used (>90%) was used to pay for the paper logbooks used at AMOLF and ARCNL. So far, a total of 15 k€ has been spent, with the first costs registered on 14/10/2019.

For the coming period we do not foresee any additional budget requirements specific to data management. The current budget provides room to explore additional tools to implement, e.g., training, subversion, and Electronic Logbooks. A crucial component will be for storage, however the budget for additional storage and backup solutions for data falls outside the budget provided for data management.

Data management team composition

The current data management team is the result of a merger of the 3 original teams that focused on 1) DM plans, logbooks and training, 2) storage, acquisition, meta-data and processing tools, and 3) replication packages and open access. The team now consists mostly of group leaders (research and support) from AMOLF and ARCNL, and the librarian who takes up many of the registration tasks (e.g., for handing out and archiving logbooks).

With most of the original data management policy implemented, the next period will focus more on training new staff, monitoring the implementation and making most data available openly. This will require some additional measures to be implemented, but mostly to evaluate how the implemented tools are being used and to identify improvements that need to be made to the implementation for effective data management.

As a result, the focus of the team is shifting, which should also be reflected in the composition of the data management team. The team should be positioned around researchers and support staff that work daily with the implemented data management tools, or work on implementation of the tools. Moreover, it should cover the expertise centers (or themes), to ensure that lines between researchers and the data management team remains short. Required expertise from support department (e.g., ICT or software) can then be requested on demand. The proposed team that will start (ensuring a slow enough transition to transfer current knowledge) in 2024 consists of:

- Group leader AMOLF and ARCNL, of which one will act as the chair
- Librarian (works with the implementation of Data Management on a daily basis, e.g. handing out and archiving logbooks)
- ICT expert (preferably the same person who is implementing ICT tools for data management)
- 3-5 junior researchers (PhD or PostDoc) representing the broad background of AMOLF and ARCNL.

You may notice that the concept of an internal "data steward" (note that we do have a data steward within the Local DCC, see section 3B) is an apparent 'blind spot' in the policy document, as the policy is written from the viewpoint of the status quo in which AMOLF-ARCNL has a Data Management team, and not a dedicated data steward. At the same time, we note that the librarian *de facto* already has some tasks that are commonly part of the package of tasks of a "Data steward", and similarly ICT takes some such roles, yet many of the tasks are also taken up by the Data Management team. We recommend that we keep reflecting on how much workload these tasks encompass and ensure that there is sufficient time reserved from personnel (e.g., librarian and ICT) to implement existing policy.