Nikhef Research Data Management Policy

The Dutch National Institute for Sub-atomic Physics Nikhef, via its mission and through the programmes, projects, and collaborations that it operates and subscribes to, is a significant producer of scientific research data, and transfer of this knowledge to third parties, i.e., industry, civil society and general public, is an integral part of Nikhef's mission. Nikhef is committed to ensuring careful management and optimal exploitation of the research data, both in the short term and the long term, in alignment with the principles on data management of NWO, and in accordance with this Policy[1].

Scope

This Policy applies to all research data that are relevant for re-use and produced as a result of Nikhef *Research Activities*, i.e.,

- all approved granted research programmes and granted research projects, and
- research projects so designated and approved by the Nikhef director, and
- any activity that results in *Published* data as per the General Principles.

This Policy shall apply without prejudice to provisions set forth in more specific agreements between Nikhef and any third party, which in all cases take precedence.

This Policy does not apply to

- data resulting from or relating to work carried out by Nikhef under contract, or under a service level agreement with other organisations, or data arising from commercial or third-party use of Nikhef facilities and installations that are not also part of *Research Activities*. Policy regarding such data is the responsibility of the contracting organisation.
- to research outputs and publications arising from Nikhef funding to third parties, to which however the principles of the NWO Open Science policy should remain applicable.
- software as a form of data in its own right (as distinct from software required to make use of the data)
- physical (collections of) items. Where any physical apparatus is unique and of essential nature for the reproducibility of research, its design drawings and properties shall be considered part of the research data to which the Policy applies, and those shall be considered a fully adequate representation thereof
- data that are personal data according to Regulation (EU) 2016/679 "GDPR", for which in all cases a specific Data Management Plan must be provided and separately approved by the Nikhef director
- data which are purely administrative in nature

[1] This document follows RFC 2119 terminology to signify the status of requirements, even when such statement are not capitalized. See https://www.ietf.org/rfc/rfc2119

General Principles

1. Nikhef *Policy* aligns with the NWO Open Science policy[2] and implements the NWO Institutes Data Management Policy Framework.
2. Both *Policy* and practice must be consistent with relevant Dutch and international legislation.
3. For the purposes of this *Policy*, the term '*Data*' refers to
    a. '*Raw*' research data directly arising as a result of experiments, measurements, and observations;
    b. '*Derived*' data which has been subject to some form of standard or automated data processing procedure, e.g. to reduce the data volume or to transform to a physically meaningful coordinate system, where such derivation cannot be reasonably reproduced;
    c. '*Published*' data, i.e. that data which is displayed or otherwise referred to in a publication (or pre-print) and based on which the scientific conclusions are derived;
    d. '*Log*' data, i.e. those data, descriptions, notes, configurations, settings, and documents that are necessary and sufficient for reproduction of the '*Published*' data.
4. Nikhef is not responsible for the use made of *Data*, except that made by its own employees.
5. *Data Management Plans* should exist for all *Data* within the scope of the *Policy*. These should be prepared in consultation with relevant stakeholders[3], before commencement of the *Research Activities*. Where no specific *Data Management Plan* exists, projects must at least implement the Guidelines and Best Practice described herein.
6. The *Research Activity* lead, so designated by the Nikhef director, is responsible for the *Data Management Plan* for the activity, or its conformance to the Guidelines and Best Practice, and must designate a person or persons responsible for providing the replication package(s) that will be deposited in the repository or repositories. *Published* data should have an associated replication package.
7. Proposals for grant funding, for those projects which result in the production or collection of *Data*, should include an abbreviated proposal of their *Data Management Plan*.
8. Where Nikhef is a subscribing partner to an external organisation or collaboration, e.g. as a member of CERN, it will seek to ensure that such organisation has a data management policy.

[2] https://www.nwo.nl/en/policies/open+science
[3] Plans should aim to streamline activities utilising existing skills and capabilities, in particular for smaller projects.

9. *Data Management Plans* should follow relevant national and international recommendations for best practice, such as DPHEP[4] or the NWO Data Management Plan form.
10. *Data* resulting from publicly funded research should be made publicly available after a limited period, unless there are specific reasons (e.g. legislation, ethical, privacy and security) why this should not happen. The length of any proprietary period should be specified in the *Data Management Plan* and justified, for example, by the reasonable needs of the research team to have a first opportunity to exploit the results of their research, including any Intellectual Property arising. In absence of specific reasons, where there are accepted norms within a scientific field or for a specific archive they should be followed.
11. The *Data* should be made available in a format and manner that allows appropriately qualified and trained researchers in the research domain to replicate the published results.
12. '*Published'* data should generally be made available within six months of the date of the relevant publication.
13. To enable *Data* to be discoverable and effectively re-used by others, sufficient metadata should be recorded and made openly available to enable other qualified researchers to understand the research and re-use potential of the data. *Published* results should include information on how to access the supporting data.

Guidelines and Best Practice

Nikhef recommends that a specific *Data Management Plan* is formulated following the guidance provided by NWO[5] and taking appropriately into account the guidelines and best practices described below. Otherwise, these guidelines and best practices represent the baseline for *Data Management* for the *Research Activity*.

1. The *Data Management Plan* or *Research Activity* lead must designate a person or persons responsible for providing the replication package(s) that will be deposited in the repository or repositories.
2. Nikhef would normally expect, upon completion of a research project, programme, or significant phase thereof, the resulting *Data* to be managed through an independently-managed domain-specific repository, a general purpose international repository such as Zenodo[6], a national repository, or a so-designated institutional repository for which long-term sustainability is ensured. The repository(ies) should be chosen so as to maximise the scientific value obtained from aggregation of related data. It may be appropriate to use different repositories for data from different stages of a study.

[4] http://www.dphep.org
[5] https://www.nwo.nl/en/documents/nwo/data-management/data-management-plan-form
[6] https://zenodo.org/

3. On deposition of *Data* in a domain-specific, international, or national repository, unique identifiers must be assigned, preferably a DOI, by the researcher or the repository, and the author(s) unambiguously identified with their affiliations, preferably with their ORCID, so that the intellectual contributions of researchers can be acknowledged. Where data is not (yet) fully Open Access, the terms and conditions of access shall be clearly indicated.

4. The deposited *Data* or replication package(s) shall be structured so that *Data* and associated metadata are self-contained and can be referenced by a collective identifier. A *Research Activity* can result in multiple *Data* structures or replication packages, which may be deposited in distinct repositories, based on the nature of the *Data* or the way of publication of the results derived therefrom.
   Without prejudice to the requirements of any repository, the *Data* structure or replication package need not be a single object or archive, but may be contained in a structured hierarchy of objects and permanently resolvable references[7]. An appropriate format or structure, such as but not limited to HEPData[8], should be chosen to ensure appropriate metadata, including the Dublin Core Metadata Element Set[9], is registered.

5. Designs, drawings, and models of experimental apparatus used in the research project are considered sufficient substitute for any physical apparatus. Such designs, drawings, and models in digital form may be preserved in an institutional repository only, and in a representation and format appropriate at the time, even if such a format is proprietary. If a description of the apparatus is published in a manner that permits re-creation by a qualified engineer, the original designs, drawings, and models need not be published.

6. During execution of the research, *Data* must be managed through and maintained on resources that ensure durability, persistency, and continuity of access. This would normally mean those resources provided centrally by, or by way of, the Nikhef institutional information technology services, or on resources so designated by the research collaboration. The choice of storage quality of service[10] must be commensurate with the value of the data.

7. *Data* that are not 'born digitally', such as (written) '*Log*' data, must be put in digital form without undue delay and curated alongside any digital research data.

---

[7] While an format resulting in a single 'file' (e.g. a *tar* archive of Root files, associated scripts, references to on-line notebooks, and a 'readme' file) may be preferred or required for deposition in a public repository, especially during execution of the research or for ease of re-use such a set of Root files, scripts, git urls, &c, could better be maintained in a hierarchical directory structure on Nikhef-provided and bit-curated storage resources. The URL to such a hierarchy on persistent and curated storage is considered a valid unique identifier for internal use if such a path is non-reassigned (e.g. is dated or programmatically generated)
[8] https://hepdata.net/submission
[9] http://dublincore.org/documents/dces/
[10] Nikhef central services provide storage of transient files ("data"), for results that can be replicated (distributed mass storage), persistent reliable storage ("project"), version-management systems, and archival (dark) storage, as well as means to make (mobile) storage suitable for results that can be replicated ("backup" facilities).

8. Plans should provide suitable quality assurance concerning the extent to which *Data* can be or have been modified. Where specific data sets are not to be retained, the processes for obtaining such data sets should be specified and conform to the standard accepted procedures within the scientific field at that time.
9. Plans may reference the general policy(ies) for the chosen repository(ies) and only include further details related to the specific project. It is the responsibility of the person preparing the data management plan to ensure that the repository policy is appropriate. Where *Data* are not to be managed through an established repository, the *Data Management Plan* will need to be more extensive and to provide reassurance on the likely stability and longevity of any repository proposed.
10. Plans should cover all *Data* expected to be produced as a result of a project or activity, and that are relevant for re-use or reproducibility, from '*Raw*' to '*Published*'.
11. Plans should specify which data are to be deposited in a repository, where and for how long, with appropriate justification. Where specific *data* are not deposited, sufficient information to re-derive such *data* must be deposited. Similarly, where re-creation of derived *data* based on deposited *data* would result in significant resource savings, the source *data* and the derivation method should be stored. The good practice criteria assume that this data is accompanied by sufficient metadata to enable re-use.
12. It is recognised that a balance may be required between the cost of data curation (e.g. for very large data sets) and the potential long term value of that data. Wherever possible Nikhef would expect the original data (i.e. from which other related data can in principle be derived) to be retained for the longest possible period, with ten years after the end of the project being a reasonable minimum. For data that by their nature cannot be re-measured, effort should be made to retain them 'in perpetuity'.

Acknowledgements

We highly appreciate the inspiration provided by the STFC Scientific Data Policy[11].

[11] http://www.stfc.ac.uk/about-us/our-purpose-and-priorities/freedom-of-information/scientific-data-policy/