Technical documentation

D2.5 Data Management Plan

Versions

Version	Date	Comment	Responsible
0.5	2020-11-03	Added list of data centres being harvested and that NORMAP is integrated as a data collection.	Øystein Godøy
0.4	2017-03-05	Added comments from Markus.	Markus Fiebig Øystein Godøy
0.3	2016-11-24	Added references to FAIR data management and licenses.	Torill Hamre Øystein Godøy
0.2	2016-10-29	Minor modifications, corrections of some sentences, addition of references.	Øystein Godøy
0.1	2016-09-02	Restructured following NSF and other templates.	Øystein Godøy
0.0	2015-11-28	Intial version	Stein Tronstad Øystein Godøy Helge Sagen

Table of Contents

Index of Tables

1 Introduction

1.1 Background

The NorDataNet project aims at *providing the scientific community with an integrated, cost-efficient and sustainable infrastructure following established standards for data documentation, archiving, search and exchange*. Focus is on coordination of efforts and requirements, enhancing interoperability and user interfaces to existing facilities, as well as development of toolboxes for data documentation in order to reduce technical and governance obstacles. The infrastructure must provide online access to data and facilitate long term preservation, in order to maximise the benefit of public funds invested in the datasets. The intention is to link existing institutional and discipline specific systems to promote science regardless of geographical and institutional location.

NorDataNet data management will be based on the FAIR Guiding Principles for Data Management[1][2], and the fundamental principles laid down in the data policies of the ICSU World Data System and the Global Earth Observation System of Systems (GEOSS):

- There will be full and open exchange of data, metadata and products shared within the system;
- All shared data, metadata and products will be made available with minimum time delay and at minimum cost;
- All shared data, metadata and products being free of charge or no more than cost of reproduction will be encouraged for research and education.

Furthermore, all data managed by the NorDataNet data centres and served through the NorDataNet portals should be preserved indefinitely[3] for future, responsible reuse, and managed under public, open licences. This imposes further requirements, of which the following are recognised by key international bodies coordinating scientific research and observation programmes, such as SCAR, IASC, SAON, WMO, IODE, GBIF, SIOS, ACTRIS, ICOS, IAGOS, Euro-ARGO and SeaDataNet:

- Ethically open data, meaning that all data be kept openly available unless concerns of privacy or indigenous rights apply, or if open data access entails risks to vulnerable sites or species.
- All data should be tagged with persistent and universally unique identifiers, so as to allow unequivocal source identification and dataset citation in publications. Data citation, in accordance with the norms and guidelines of DataCite[4] and FORCE11[5], aids reproducibility of research and formal recognition of the dataset originators.

---

1   Wilkinson et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship Scientific Data 3, Article number: 160018 (2016).
2   http://ec.europa.eu/research/participants/data/ref/h2020/.../h2020-hi-oa-data-mgt_en.pdf
3   At least for observational data that is not reproducible.
4   https://www.datacite.org/services/cite-your-data.html
5   Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014

Good data management implies a number of activities which ensure that data are discoverable and accessible, that they may be understood and used, and that they are looked after in the long term. The participating data centres should be certified as operating in accordance with specific standards, best practices and requirements for safe, long-term data storage. Accreditation of "trusted repositories" is provided by several international bodies, including ISO, WMO, IODE, and WDS. NorDataNet will follow the activities within Research Data Alliance Repository Audit and Certification DSA–WDS Partnership WG to base its recommendation on this work.

1.2 Scope

Norwegian Scientific Data Network (NorDataNet) is funded by the Research Council of Norway under grant contract 245967. NorDataNet is focusing on integrating existing and new data centres into a unified infrastructure supporting scientific data management. NorDataNet does not generate data, but is a service for scientists and projects who do.

NorDataNet will be providing solutions for scientific data originating in the natural sciences, and serve the data by means of a shared, geoscientific data catalogue and common data interfaces. In the future, solutions may be expanded to accommodate data from other scientific disciplines, most likely by inviting new partner institutions, or in collaborative projects with other data networks.

The submission, archival, management and publication of individual datasets will be handled by each partner data centre, by their own systems and standards. Common protocols, formats and interfaces will be implemented by NorDataNet, for the purpose of data sharing, dissemination, and integration.

1.3 Audience

This document is written for both internal and external audiences of the project. This includes, but is not limited to data consumers, data providers, data managers, system developers, and software architects. The first two groups include scientists. It is part of the foundation for the design process but more important it is providing background information for potential data providers and consumers considering to use NorDataNet.

1.4 Applicable documents

[1] Norwegian Scientific Data Network (NorDataNet) Project Plan, revision 19, NFR Contract 245967
[2] NorDataNet Concepts and acronyms
[3] NorDataNet Use cases
[4] NorDataNet System requirements
[5] NorDataNet System design and interfaces
[6] NorDataNet System implementation and integration plan
[7] NorDataNet Risk Management Plan
[8] IASC Data Policy, April 16, 2013.

2 Administrative information

The principal investigators are listed below.

Institution	Contact	Role	Email
Norwegian Meteorological Institute	Øystein Godøy	Project Coordinator Lead WP 2	o.godoy@met.no
Institute of Marine Research	Helge Sagen	Lead WP 3	h.sagen@imr.no
NERSC	Torill Hamre	Lead WP 5	torill.hamre@nersc.no
NILU	Markus Fiebig	Lead WP 4	markus.fiebig@nilu.no
Norwegian Polar Institute	Stein Tronstad	Lead WP 1	stein.tronstad@npolar.no

The main point of contact is Øystein Godøy (+47 9802 4433). Updates to the document is indicated in the versions table.

3 Data collection

3.1 Background

NorDataNet is supporting an open data space[6] and relies on data hosted by existing or future data management systems. This covers a wide range of data types and formats. Initially this is referring to geoscientific data for proof of concept, but will evolve into an interdisciplinary system.

Volumes, formats and other specific information on the data served are not known, but are more an issue for the contributing data centres than the integration of these through NorDataNet. NorDataNet relies on a number of standardised interfaces to metadata and data. This is necessary in order to facilitate integration of metadata and data across data centres. In order to simplify the process for contributing data centres NorDataNet will recommend some file formats for data submission and exchange. This is however a deliverable within the project and this document will be when this deliverable is available along with guidance material.

For access to data, usage of the OPeNDAP protocol is preferred due to its relation to the generic data model Common Data Model defined by UNIDATA. It is however acknowledged that other solutions may have to be evaluated in order to support the open data space.

NorDataNet will not reinvent methodology, but always attempt to reuse existing methodology, terminology and software where possible. In this process NorDataNet will contribute to ongoing standardisation efforts on regional and global scales.

---

6   In this context an open data space implies that any type of data (not only geoscientific) can be handled by the system. This improves the support for interdisciplinary science. Furthermore, there is not an upper threshold of the number of datasets that can be handled.

3.2 Types of data

Initially this is categorised as gridded, profile, time series and trajectory data. The origin of the data may be instrumental (e.g. temperature) or manual (e.g. cloud types) observations, or higher order processed data produced using statistical or numerical simulation methods.

3.3 Data centres integrated

Discovery metadata from the following data centres are harvested regularly into the NorDataNet data catalogue at the national level. Work is in progress to extend this list of data centres.

Data centre	URL	Contact	Comment
Norwegian Marine Data Centre	http://www.nmdc.no/	Helge Sagen	Subsystem is currently under development. It integrates information from many partners. Discovery metadata is served through OAI-PMH, the interoperability at the data level is varying from FTP with no standardisation to OPeNDAP and NetCDF/CF.
Norwegian Meteorological Institute	http://adc.met.no/	Øystein Godøy	This subsystem is integrated through NorDataNet. Discovery metadata interfaces are available, data interoperability is supported using OGC WMS and OPeNDAP. Will integrate relevant data from WMO GTS and NBS (Sentinel).
Norwegian Polar Institute	http://data.npolar.no/	Stein Tronstad	Metadata interoperability interfaces are available. Some challenges for data interoperability.
NERSC		Torill Hamre	Metadata interoperability interfaces are available and interoperability at the data level is also available.
EBAS	http://ebas.nilu.no/	Markus Fiebig	Metadata interoperability interfaces are available and interoperability at the data level is also available.

In addition information in the Norwegian Area of Interest from a number of international data centres

are also harvested. Among these are PANGAEA, World Glacier Monitoring Service, National Snow and Ice Data Centre, Polish Polar Data Centre, British Antarctic Survey, National polar Institute (Japan), Polar Research Institute of China, Environment Canada, and WSL/SLF in Switzerland. Work is in progress to extend this list and to improve the quality of harvested metadata.

3.4 Data collections integrated

NorDataNet has integrated the data collection of the Norwegian Satellite Earth Observation Database for Marine and Polar Research (NORMAP) as a separate data collection in the NorDataNet system.

4 Ethics and Legal compliance

4.1 Legal aspects

Data ownership will remain with the originating institutes and scientists. In accordance with national and international norms for data publishing, NorDataNet recommends that data are published under an open public licence such as the Creative Commons attribution licence (CC BY)[7] or Norwegian License for Open Government Data (NLOD)[8]. Exceptions from this principle must be justified and will normally be related to privacy concerns or environmental protection issues. Acknowledging the role of participating frameworks in setting data access policies, NorDataNet partners will work towards adopting open public data licenses in frameworks where this isn't the case yet. In the initial setup NorDataNet is not handling sensitive data, but the design will include this perspective.

NorDataNet will develop guidance documentation for contributing data centres and include continuous monitoring of this issue in the context of the projects risk management system [7] .

4.2 Policies for access, sharing and reuse

NorDataNet will promote free and open access to data in accordance with OECD data sharing principles. There are data or contexts where this is impossible. Such situations has to be justified referring to privacy or security issues. NorDataNet promotes the concept of ethically open access according to the IASC Data Policy [8] . According to this, the only exceptions to this requirement of full, free, and open access are (quoted):

- where human subjects are involved, confidentiality shall be protected as appropriate and guided by the principles of informed consent;

- where local and traditional knowledge is concerned, rights of the knowledge holders shall not be compromised;

- where data release may cause harm, specific aspects of the data may need to be kept protected (for example, locations of nests of endangered birds or locations of sacred sites).

While sharing data, transparency and traceability is important. NorDataNet has activities focusing on implementation and best practises of utilisation of persistent identifiers in this context. All partners

---

7   https://creativecommons.org/licenses/by/4.0/

8   http://data.norge.no/nlod/en

have established a dialogue with BIBSYS concerning integration of Document Object Identifiers (DOIs), but effective utilisation of such persistent identifiers, both for final and pre-final data products, while sharing and reusing data is an issue to be examined in the project.

Unless specifically required for individual datasets, data access or data usage will not be constrained. Institutes and science groups will, however, have an option to limit data access for a limited period of time, to protect scientific first rights. Such embargo periods should normally not exceed 1 year, at any case not extend longer than 2 years. This duration is to be discussed with stakeholders.

5 Data documentation and sharing

5.1 Introduction

NorDataNet will accommodate highly diverse data from many scientific disciplines, for which no universal documentation of sharing standards exist. It is expected that international de jure or de facto standards are adhered to whenever applicable.

All data should be made available in non-proprietary standards, and be readable with open source or openly documented software.

5.2 Metadata

NorDataNet require at least three levels of metadata. Discovery metadata is required to find relevant data, use metadata to understand the data found and configuration metadata are required by the data management system to effectively serve data to the users.

All data served by NorDataNet must be fully and adequately documented for independent, future reuse. In most cases this requires both a general data set description in the form of "discovery level" metadata, and more specific descriptions of instruments, operational parameters, field procedures etc in a "use metadata" document following the dataset. Discovery level metadata must be compliant with the DIF and ISO 19115 metadata standards as these are widely used within relevant communities. Challenges remain concerning controlled vocabularies. Discovery metadata will be harvested using OAI-PMH and use metadata must follow the data leaving datasets self explainable in a standardised manner (e.g. as supported by NetCDF/CF).

In the context of metadata, more important than the standard itself is the terminology used within the standard. In order to effectively map between disciplines and data management frameworks, NorDataNet require machine readable controlled vocabularies. This implies information provided as Resource Description Framework (RDF) or preferably Simple Knowledge Organisation System (SKOS).

5.3 Data

All data in NorDataNet will be openly available on the Internet, either in the form of downloadable files, or served directly by machine-accessible digital services (e.g. combining datasets). User friendly, robust and well known protocols like HTTP, FTP, and OPeNDAP will be employed. Web services that are becoming standards in international communities will be supported when relevant. All data

will be searchable through the NorDataNet portals and through international data portals such as GCMD, WIS, SeaDataNet, and others.

In order to facilitate combination of datasets into larger datasets meeting user requirements from various communities, integration of data need to be an automated process. This calls for application of a common data model and machine readable controlled terminologies. Initially NetCDF following the Climate and convention will be supported, while evaluating the usability of others in the process. Unless such standardisation is achieved, higher order services and translation of information between disciplines/user communities is neither possible nor sustainable.

5.4 Quality control.

Quality control requirements will vary between disciplines and projects, and are not specified by NorDataNet. The quality control procedures applied should, however be adequately described in the metadata implying that all datasets should as a minimum have a general statement in discovery metadata on the quality of the dataset, preferably accompanied with quality information on individual data values. How to achieve this is a deliverable of the project and must follow ongoing activities within GEO, RDA and ICSU.

6 Long term data preservation

6.1 Infrastructure

Long term data storage, backup and management will remain the responsibility of individual, participating data centres, or by national government funded data archives, notably the Norwegian Infrastructure for research Data (NIRD). In general NorDataNet will utilise national resources like NIRD wherever possible and development of local systems will be coordinated with NIRD following a requirement analysis. This implies that investments in local system are a result of a gap analysis and a dialogue with NIRD leaving this as the best solution.

In this context NorDataNet will work with NIRD on developing a sustainable business model.

6.2 Life cycles

Each data centre must establish life cycle management for datasets. NorDataNet will develop guidelines and interfaces. Life cycle management for data establishes policies for the flow of data, from its creation through active usage to the time where it potentially becomes obsolete. Details on the life cycles will be developed in dialogue with the user community. As a general rule, observations are preserved forever, but not necessarily in the original format. Transformations are however required to be lossless and tools to provide the original format is required at least on a temporal basis. Analysed data, whether generated by statistical or numerical methods have a minimum preservation time of 10 years.